Generalized score matching for non-negative data

Yu, Shiqing; Drton, Mathias; Shojaie, Ali

Published in:
Journal of Machine Learning Research

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Journal of Machine Learning Research 20 (2019) 1-33 Submitted 5/18; Revised 2/19; Published 4/19

Nonuniformity of P-values Can Occur Early in Diverging Dimensions

Yingying Fan FANYINGY@MARSHALL.USC.EDU
Data Sciences and Operations Department
University of Southern California
Los Angeles, CA 90089, USA

Emre Demirkaya DEMIRKAY@USC.EDU
Business Analytics & Statistics
The University of Tennessee, Knoxville
Knoxville, TN 37996-4140, USA

Jinchi Lv JINCHILV@MARSHALL.USC.EDU
Data Sciences and Operations Department
University of Southern California
Los Angeles, CA 90089, USA

Editor: Sara van de Geer

Abstract

Evaluating the joint significance of covariates is of fundamental importance in a wide range of applications. To this end, p-values are frequently employed and produced by algorithms that are powered by classical large-sample asymptotic theory. It is well known that the conventional p-values in Gaussian linear model are valid even when the dimensionality is a non-vanishing fraction of the sample size, but can break down when the design matrix becomes singular in higher dimensions or when the error distribution deviates from Gaussianity. A natural question is when the conventional p-values in generalized linear models become invalid in diverging dimensions. We establish that such a breakdown can occur early in nonlinear models. Our theoretical characterizations are confirmed by simulation studies.

Keywords: Nonuniformity, p-value, breakdown point, generalized linear model, high dimensionality, joint significance testing

1. Introduction

In many applications it is often desirable to evaluate the significance of covariates in a predictive model for some response of interest. Identifying a set of significant covariates can facilitate domain experts to further probe their causal relationships with the response. Ruling out insignificant covariates can also help reduce the fraction of false discoveries and narrow down the scope of follow-up experimental studies by scientists. These tasks certainly require an accurate measure of feature significance in finite samples. The tool of p-values has provided a powerful framework for such investigations.

As p-values are routinely produced by algorithms, practitioners should perhaps be aware that those p-values are usually based on classical large-sample asymptotic theory. For ex-

ample, marginal p-values have been employed frequently in large-scale applications when the number of covariates p greatly exceeds the number of observations n. Those p-values are based on marginal regression models linking each individual covariate to the response separately. In these marginal regression models, the ratio of sample size to model dimensionality is equal to n, which results in justified p-values as sample size increases. Yet due to the correlations among the covariates, we often would like to investigate the joint significance of a covariate in a regression model conditional on all other covariates, which is the main focus of this paper. A natural question is whether conventional joint p-values continue to be valid in the regime of diverging dimensionality p.

It is well known that fitting the linear regression model with $p > n$ using the ordinary least squares can lead to perfect fit giving rise to zero residual vector, which renders the p-values undefined. When $p \leq n$ and the design matrix is nonsingular, the p-values in the linear regression model are well defined and valid thanks to the exact normality of the least-squares estimator when the random error is Gaussian and the design matrix is deterministic. When the error is non-Gaussian, Huber (1973) showed that the least-squares estimator can still be asymptotically normal under the assumption of $p = o(n)$, but is generally no longer normal when $p = o(n)$ fails to hold, making the conventional p-values inaccurate in higher dimensions. For the asymptotic properties of M-estimators for robust regression, see, for example, Huber (1973); Portnoy (1984, 1985) for the case of diverging dimensionality $p = o(n)$ and Karoui et al. (2013); Bean et al. (2013) for the scenario when the dimensionality p grows proportionally to sample size n.

We have seen that the conventional p-values for the least-squares estimator in linear regression model can start behaving wildly and become invalid when the dimensionality p is of the same order as sample size n and the error distribution deviates from Gaussianity. A natural question is whether similar phenomenon holds for the conventional p-values for the maximum likelihood estimator (MLE) in the setting of diverging-dimensional nonlinear models. More specifically, we aim to answer the question of whether $p \sim n$ is still the breakdown point of the conventional p-values when we move away from the regime of linear regression model, where $\sim$ stands for asymptotic order. To simplify the technical presentation, in this paper we adopt the generalized linear model (GLM) as a specific family of nonlinear models (McCullagh and Nelder, 1989). The GLM with a canonical link assumes that the conditional distribution of $\mathbf{y}$ given $\mathbf{X}$ belongs to the canonical exponential family, having the following density function with respect to some fixed measure

$$f_n(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) \equiv \prod_{i=1}^{n} f_0(y_i; \theta_i) = \prod_{i=1}^{n} \left\{ c(y_i) \exp\left[\frac{y_i \theta_i - b(\theta_i)}{\phi} \right] \right\}, \tag{1}$$

where $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ is an $n \times p$ design matrix with $\mathbf{x}_j = (x_{1j}, \cdots, x_{nj})^T$, $j = 1, \cdots, p$, $\mathbf{y} = (y_1, \cdots, y_n)^T$ is an n-dimensional response vector, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a p-dimensional regression coefficient vector, $\{f_0(y; \theta) : \theta \in \mathbb{R}\}$ is a family of distributions in the regular exponential family with dispersion parameter $\phi \in (0, \infty)$, and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T = \mathbf{X}\boldsymbol{\beta}$. As is common in GLM, the function $b(\theta)$ in (1) is implicitly assumed to be twice continuously differentiable with $b''(\theta)$ always positive. Popularly used GLMs include the linear regression model, logistic regression model, and Poisson regression model for continuous, binary, and count data of responses, respectively.

The key innovation of our paper is the formal justification that the conventional p-values in nonlinear models of GLMs can become invalid in diverging dimensions and such a breakdown can occur *much earlier* than in linear models, which spells out a fundamental difference between linear models and nonlinear models. To begin investigating p-values in diverging-dimensional GLMs, let us gain some insights into this problem by looking at the specific case of logistic regression. Recently, Candès (2016) established an interesting phase transition phenomenon of perfect hyperplane separation for high-dimensional classification with an elegant probabilistic argument. Suppose we are given a random design matrix $\mathbf{X} \sim N(\mathbf{0}, I_n \otimes I_p)$ and arbitrary binary y_i's that are not all the same. The phase transition of perfect hyperplane separation happens at the point $p/n = 1/2$. With such a separating hyperplane, there exist some $\boldsymbol{\beta}^* \in \mathbb{R}^p$ and $t \in \mathbb{R}$ such that $\mathbf{x}_i^T \boldsymbol{\beta}^* > t$ for all cases $y_i = 1$ and $\mathbf{x}_i^T \boldsymbol{\beta}^* < t$ for all controls $y_i = 0$. Let us fit a logistic regression model with an intercept. It is easy to show that multiplying the vector $(-t, (\boldsymbol{\beta}^*)^T)^T$ by a divergence sequence of positive numbers c, we can obtain a sequence of logistic regression fits with the fitted response vector approaching $\mathbf{y} = (y_1, \cdots, y_n)^T$ as $c \to \infty$. As a consequence, the MLE algorithm can return a pretty wild estimate that is close to infinity in topology when the algorithm is set to stop. Clearly, in such a case the p-value of the MLE is no longer justified and meaningful. The results in Candès (2016) have two important implications. First, such results reveal that unlike in linear models, p-values in nonlinear models can break down and behave wildly when p/n is of order $1/2$; see Karoui et al. (2013); Bean et al. (2013) and discussions below. Second, these results motivate us to characterize the breakdown point of p-values in nonlinear GLMs with $p \sim n^{\alpha_0}$ in the regime of $\alpha_0 \in [0, 1/2)$. In fact, our results show that the breakdown point can be even much earlier than $n/2$.

It is worth mentioning that our work is different in goals from the limited but growing literature on p-values for high-dimensional nonlinear models, and makes novel contributions to such a problem. The key distinction is that existing work has focused primarily on identifying the scenarios in which conventional p-values or their modifications continue to be valid with some sparsity assumption limiting the growth of intrinsic dimensions. For example, Fan and Peng (2004) established the oracle property including the asymptotic normality for nonconcave penalized likelihood estimators in the scenario of $p = o(n^{1/5})$, while Fan and Lv (2011) extended their results to the GLM setting of non-polynomial (NP) dimensionality. In the latter work, the p-values were proved to be valid under the assumption that the intrinsic dimensionality $s = o(n^{1/3})$. More recent work on high-dimensional inference in nonlinear model settings includes van de Geer et al. (2014); Athey et al. (2016) under sparsity assumptions. In addition, two tests were introduced in Guo and Chen (2016) for high-dimensional GLMs without or with nuisance regression parameters, but the p-values were obtained for testing the global hypothesis for a given set of covariates, which is different from our goal of testing the significance of individual covariates simultaneously. Portnoy (1988) studied the asymptotic behavior of the MLE for exponential families under the classical i.i.d. non-regression setting, but with diverging dimensionality. In contrast, our work under the GLM assumes the regression setting in which the design matrix $\mathbf{X}$ plays an important role in the asymptotic behavior of the MLE $\widehat{\boldsymbol{\beta}}$. The validity of the asymptotic normality of the MLE was established in Portnoy (1988) under the condition of $p = o(n^{1/2})$, but the precise breakdown point in diverging dimensionality was not investigated therein. Another line of work is focused on generating asymptotically valid p-values when p/n converges to a

fixed positive constant. For instance, Karoui et al. (2013) and Bean et al. (2013) considered M-estimators in the linear model and showed that their variance is greater than classically predicted. Based on this result, it is possible to produce p-values by making adjustments for the inflated variance in high dimensions. Recently, Sur and Candès (2018) showed that similar adjustment is possible for the likelihood ratio test (LRT) for logistic regression. Our work differs from this line of work in two important aspects. First, our focus is on the *classical* p-values and their validity. Second, their results concern dimensionality that is comparable to sample size, while we aim to analyze the problem for a lower range of dimensionality and pinpoint the exact breakdown point of p-values.

The rest of the paper is organized as follows. Section 2 provides characterizations of p-values in low dimensions. We establish the nonuniformity of GLM p-values in diverging dimensions in Section 3. Section 4 presents several simulation examples verifying the theoretical phenomenon. We discuss some implications of our results in Section 5. The proofs of all the results are relegated to the Appendix.

2. Characterizations of P-values in Low Dimensions

To pinpoint the breakdown point of GLM p-values in diverging dimensions, we start with characterizing p-values in low dimensions. In contrast to existing work on the asymptotic distribution of the penalized MLE, our results in this section focus on the asymptotic normality of the unpenalized MLE in diverging-dimensional GLMs, which justifies the validity of conventional p-values. Although Theorems 1 and 4 to be presented in Sections 2.2 and A are in the conventional sense of relatively small p, to the best of our knowledge such results are not available in the literature before in terms of the maximum range of dimensionality p without any sparsity assumption.

2.1. Maximum likelihood estimation

For the GLM (1), the log-likelihood $\log f_n(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta})$ of the sample is given, up to an affine transformation, by

$$\ell_n(\boldsymbol{\beta}) = n^{-1} \left[\mathbf{y}^T \mathbf{X}\boldsymbol{\beta} - \mathbf{1}^T \mathbf{b}(\mathbf{X}\boldsymbol{\beta}) \right], \tag{2}$$

where $\mathbf{b}(\boldsymbol{\theta}) = (b(\theta_1), \cdots, b(\theta_n))^T$ for $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T \in \mathbb{R}^n$. Denote by $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T \in \mathbb{R}^p$ the MLE which is the maximizer of (2), and

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = (b'(\theta_1), \cdots, b'(\theta_n))^T \quad \text{and} \quad \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \operatorname{diag}\{b''(\theta_1), \cdots, b''(\theta_n)\}. \tag{3}$$

A well-known fact is that the n-dimensional response vector $\mathbf{y}$ in GLM (1) has mean vector $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariance matrix $\phi\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Clearly, the MLE $\widehat{\boldsymbol{\beta}}$ is given by the unique solution to the score equation

$$\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})] = \mathbf{0} \tag{4}$$

when the design matrix $\mathbf{X}$ is of full column rank p.

It is worth mentioning that for the linear model, the score equation (4) becomes the well-known normal equation $\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$ which admits a closed form solution. On the other hand, equation (4) does not admit a closed form solution in general nonlinear models. This fact due to the nonlinearity of the mean function $\boldsymbol{\mu}(\cdot)$ causes the key difference between

the linear and nonlinear models. In future presentations, we will occasionally use the term *nonlinear GLMs* to exclude the linear model from the family of GLMs when necessary.

We will present in the next two sections some sufficient conditions under which the asymptotic normality of MLE holds. In particular, Section 2.2 concerns the case of fixed design and Section A deals with the case of random design. In addition, Section 2.2 allows for general regression coefficient vector $\boldsymbol{\beta}_0$ and the results extend some existing ones in the literature, while Section A assumes the global null $\boldsymbol{\beta}_0 = \mathbf{0}$ and Gaussian random design which enable us to pinpoint the exact breakdown point of the asymptotic normality for the MLE.

2.2. Conventional p-values in low dimensions under fixed design

Recall that we condition on the design matrix $\mathbf{X}$ in this section. We first introduce a deviation probability bound that facilitates our technical analysis. Consider both cases of bounded responses and unbounded responses. In the latter case, assume that there exist some constants $M, v_0 > 0$ such that

$$\max_{1\le i\le n} E\left\{\exp\left[\frac{|y_i - b'(\theta_{0,i})|}{M}\right] - 1 - \frac{|y_i - b'(\theta_{0,i})|}{M}\right\} M^2 \le \frac{v_0}{2} \tag{5}$$

with $(\theta_{0,1}, \cdots, \theta_{0,n})^T = \boldsymbol{\theta}_0 = \mathbf{X}\boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0 = (\beta_{0,1}, \cdots, \beta_{0,p})^T$ denotes the true regression coefficient vector in model (1). Then by Fan and Lv (2011, 2013), it holds that for any $\mathbf{a} \in \mathbb{R}^n$,

$$P\left(\left|\mathbf{a}^T\mathbf{Y} - \mathbf{a}^T\boldsymbol{\mu}(\boldsymbol{\theta}_0)\right| > \|\mathbf{a}\|_2\,\varepsilon\right) \le \varphi(\varepsilon), \tag{6}$$

where $\varphi(\varepsilon) = 2e^{-c_1\varepsilon^2}$ with $c_1 > 0$ some constant, and $\varepsilon \in (0, \infty)$ if the responses are bounded and $\varepsilon \in (0, \|\mathbf{a}\|_2/\|\mathbf{a}\|_\infty]$ if the responses are unbounded.

For nonlinear GLMs, the MLE $\widehat{\boldsymbol{\beta}}$ solves the nonlinear score equation (4) whose solution generally does not admit an explicit form. To address such a challenge, we construct a solution to equation (4) in an asymptotically shrinking neighborhood of $\boldsymbol{\beta}_0$ that meets the MLE $\widehat{\boldsymbol{\beta}}$ thanks to the uniqueness of the solution. Specifically, define a neighborhood of $\boldsymbol{\beta}_0$ as

$$\mathcal{N}_0 = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_\infty \le n^{-\gamma}\log n\} \tag{7}$$

for some constant $\gamma \in (0, 1/2]$. Assume that $p = O(n^{\alpha_0})$ for some $\alpha_0 \in (0, \gamma)$ and let $b_n = o\{\min(n^{1/2-\gamma}\sqrt{\log n}, s_n^{-1}n^{2\gamma-\alpha_0-1/2}/(\log n)^2\}$ be a diverging sequence of positive numbers, where s_n is a sequence of positive numbers that will be specified in Theorem 1 below. We need some basic regularity conditions to establish the asymptotic normality of the MLE $\widehat{\boldsymbol{\beta}}$.

Condition 1 *The design matrix* $\mathbf{X}$ *satisfies*

$$\left\|\left[\mathbf{X}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{X}\right]^{-1}\right\|_\infty = O(b_n n^{-1}), \tag{8}$$

$$\max_{\boldsymbol{\beta}\in\mathcal{N}_0} \max_{j=1}^p \lambda_{\max}\left[\mathbf{X}^T \mathrm{diag}\left\{|\mathbf{x}_j| \circ \left|\boldsymbol{\mu}''(\mathbf{X}\boldsymbol{\beta})\right|\right\}\mathbf{X}\right] = O(n) \tag{9}$$

with $\circ$ *denoting the Hadamard product and derivatives understood componentwise. Assume that* $\max_{j=1}^p \|\mathbf{x}_j\|_\infty < c_1^{1/2}\{n/(\log n)\}^{1/2}$ *if the responses are unbounded.*

Condition 2 *The eigenvalues of* $n^{-1}\mathbf{A}_n$ *are bounded away from* 0 *and* ∞, $\sum_{i=1}^n (\mathbf{z}_i^T \mathbf{A}_n^{-1} \mathbf{z}_i)^{3/2} = o(1)$, *and* $\max_{i=1}^n E|y_i - b'(\theta_{0,i})|^3 = O(1)$, *where* $\mathbf{A}_n = \mathbf{X}^T \mathbf{\Sigma}(\boldsymbol{\theta}_0)\mathbf{X}$ *and* $(\mathbf{z}_1, \cdots, \mathbf{z}_n)^T = \mathbf{X}$.

Conditions 1 and 2 put some basic restrictions on the design matrix $\mathbf{X}$ and a moment condition on the responses. For the case of linear model, bound (8) becomes $\|(\mathbf{X}^T\mathbf{X})^{-1}\|_\infty = O(b_n/n)$ and bound (9) holds automatically since $b'''(\theta) \equiv 0$. Condition 2 is related to the Lyapunov condition.

Theorem 1 (Asymptotic normality) *Assume that Conditions 1–2 and probability bound* (6) *hold. Then*

a) there exists a unique solution $\widehat{\boldsymbol{\beta}}$ *to score equation (4) in* $\mathcal{N}_0$ *with asymptotic probability one;*

b) the MLE $\widehat{\boldsymbol{\beta}}$ *satisfies that for each vector* $\mathbf{u} \in \mathbb{R}^p$ *with* $\|\mathbf{u}\|_2 = 1$ *and* $\|\mathbf{u}\|_1 = O(s_n)$,

$$(\mathbf{u}^T \mathbf{A}_n^{-1} \mathbf{u})^{-1/2} (\mathbf{u}^T \widehat{\boldsymbol{\beta}} - \mathbf{u}^T \boldsymbol{\beta}_0) \xrightarrow{\mathscr{D}} N(0, \phi) \tag{10}$$

and specifically for each $1 \le j \le p$,

$$(\mathbf{A}_n^{-1})_{jj}^{-1/2} (\widehat{\beta}_j - \beta_{0,j}) \xrightarrow{\mathscr{D}} N(0, \phi), \tag{11}$$

where $\mathbf{A}_n = \mathbf{X}^T \mathbf{\Sigma}(\boldsymbol{\theta}_0)\mathbf{X}$ *and* $(\mathbf{A}_n^{-1})_{jj}$ *denotes the* j*th diagonal entry of matrix* $\mathbf{A}_n^{-1}$.

Theorem 1 establishes the asymptotic normality of the MLE and consequently justifies the validity of the conventional p-values in low dimensions. Note that for simplicity, we present here only the marginal asymptotic normality, and the joint asymptotic normality also holds for the projection of the MLE onto any fixed-dimensional subspace. This result can also be extended to the case of misspecified models; see, for example, Lv and Liu (2014).

As mentioned in the Introduction, the asymptotic normality was shown in Fan and Lv (2011) for nonconcave penalized MLE having intrinsic dimensionality $s = o(n^{1/3})$. In contrast, our result in Theorem 1 allows for the scenario of $p = o(n^{1/2})$ with no sparsity assumption in view of our technical conditions. In particular, we see that the conventional p-values in GLMs generally remain valid in the regime of slowly diverging dimensionality $p = o(n^{1/2})$.

3. Nonuniformity of GLM P-values in Diverging Dimensions

So far we have seen that for nonlinear GLMs, the p-values can be valid when $p = o(n^{1/2})$ as shown in Section 2, and can become meaningless when $p \ge n/2$ as discussed in the Introduction. Apparently, there is a big gap between these two regimes of growth of dimensionality p. To provide some guidance on the practical use of p-values in nonlinear GLMs, it is of crucial importance to characterize their breakdown point. To highlight the main message with simplified technical presentation, hereafter we content ourselves with the specific case of logistic regression model for binary response. Moreover, we investigate the distributional property in (11) for the scenario of true regression coefficient vector $\boldsymbol{\beta}_0 = \mathbf{0}$, that is, under the global null. We argue that this specific model is sufficient for our purpose because if the

conventional p-values derived from MLEs fail (i.e., (11) fails) for at least one $\boldsymbol{\beta}_0$ (in particular $\boldsymbol{\beta}_0 = \mathbf{0}$), then conventional p-values are not justified. Therefore, the breakdown point for logistic regression is at least the breakdown point for general nonlinear GLMs. This argument is fundamentally different from that of proving the overall validity of conventional p-values, where one needs to prove the asymptotic normality of MLEs under general GLMs rather than any specific model.

3.1. The wild side of nonlinear regime

For the logistic regression model (1), we have $b(\theta) = \log(1+e^{\theta})$, $\theta \in \mathbb{R}$ and $\phi = 1$. The mean vector $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariance matrix $\phi\boldsymbol{\Sigma}(\boldsymbol{\theta})$ of the n-dimensional response vector $\mathbf{y}$ given by (3) now take the familiar form of $\boldsymbol{\mu}(\boldsymbol{\theta}) = \left(\frac{e^{\theta_1}}{1+e^{\theta_1}}, \cdots, \frac{e^{\theta_n}}{1+e^{\theta_n}}\right)^T$ and

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \operatorname{diag}\left\{\frac{e^{\theta_1}}{(1+e^{\theta_1})^2}, \cdots, \frac{e^{\theta_n}}{(1+e^{\theta_n})^2}\right\}$$

with $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T = \mathbf{X}\boldsymbol{\beta}$. In many real applications, one would like to interpret the significance of each individual covariate produced by algorithms based on the conventional asymptotic normality of the MLE as established in Theorem 1. As argued at the beginning of this section, in order to justify the validity of p-values in GLMs, the underlying theory should at least ensure that the distributional property (11) holds for logistic regression under the global null. As we will see empirically in Section 4, as the dimensionality increases, p-values from logistic regression under the global null have a distribution that is skewed more and more toward zero. Consequently, classical hypothesis testing methods which reject the null hypothesis when p-value is less than the pre-specified level α would result in more false discoveries than the desired level of α. As a result, practitioners may simply lose the theoretical backup and the resulting decisions based on the p-values can become ineffective or even misleading. For this reason, it is important and helpful to identify the breakdown point of p-values in diverging-dimensional logistic regression model under the global null.

Characterizing the breakdown point of p-values in nonlinear GLMs is highly nontrivial and challenging. First, the nonlinearity generally causes the MLE to take no analytical form, which makes it difficult to analyze its behavior in diverging dimensions. Second, conventional probabilistic arguments for establishing the central limit theorem of MLE only enable us to see when the distributional property holds, but not exactly at what point it fails. To address these important challenges, we introduce novel geometric and probabilistic arguments presented later in the proofs of Theorems 2–3 that provide a rather delicate analysis of the MLE. In particular, our arguments unveil that the early breakdown point of p-values in nonlinear GLMs is essentially due to the *nonlinearity* of the mean function $\boldsymbol{\mu}(\cdot)$. This shows that p-values can behave wildly much early on in diverging dimensions when we move away from linear regression model to nonlinear regression models such as the widely applied logistic regression; see the Introduction for detailed discussions on the p-values in diverging-dimensional linear models.

Before presenting the main results, let us look at the specific case of logistic regression model under the global null. In such a scenario, it holds that $\boldsymbol{\theta}_0 = \mathbf{X}\boldsymbol{\beta}_0 = \mathbf{0}$ and thus

$\mathbf{\Sigma}(\boldsymbol{\theta}_0) = 4^{-1} I_n$, which results in

$$\mathbf{A}_n = \mathbf{X}^T \mathbf{\Sigma}(\boldsymbol{\theta}_0) \mathbf{X} = 4^{-1} \mathbf{X}^T \mathbf{X}.$$

In particular, we see that when $n^{-1}\mathbf{X}^T\mathbf{X}$ is close to the identity matrix I_p, the asymptotic standard deviation of the jth component $\widehat{\beta}_j$ of the MLE $\widehat{\boldsymbol{\beta}}$ is close to $2n^{-1/2}$ when the asymptotic theory in (11) holds. As mentioned in the Introduction, when $p \geq n/2$ the MLE can blow up with excessively large variance, a strong evidence against the distributional property in (11). In fact, one can also observe inflated variance of the MLE relative to what is predicted by the asymptotic theory in (11) even when the dimensionality p grows at a slower rate with sample size n. As a consequence, the conventional p-values given by algorithms according to property (11) can be much biased toward zero and thus produce more significant discoveries than the truth. Such a breakdown of conventional p-values is delineated clearly in the simulation examples presented in Section 4.

3.2. Main results

We now present the formal results on the invalidity of GLM p-values in diverging dimensions.

Theorem 2 (Uniform orthonormal design) [1] *Assume that $n^{-1/2}\mathbf{X}$ is uniformly distributed on the Stiefel manifold $V_p(\mathbb{R}^n)$ consisting of all $n \times p$ orthonormal matrices. Then for the logistic regression model under the global null, the asymptotic normality of the MLE established in (11) fails to hold when $p \sim n^{2/3}$, where $\sim$ stands for asymptotic order.*

Theorem 3 (Correlated Gaussian design) *Assume that $\mathbf{X} \sim N(\mathbf{0}, I_n \otimes \mathbf{\Sigma})$ with covariance matrix $\mathbf{\Sigma}$ nonsingular. Then for the logistic regression model under the global null, the same conclusion as in Theorem 2 holds.*

Theorem 4 in Appendix A states that under the global null in GLM with Gaussian design, the p-value based on the MLE remains valid as long as the dimensionality p diverges with n at a slower rate than $n^{2/3}$. This together with Theorems 2 and 3 shows that under the global null, the exact breakdown point for the uniformity of p-value is $n^{2/3}$. We acknowledge that these results are mainly for theoretical interests because in practice one cannot check precisely whether the global null assumption holds or not. However, these results clearly suggest that in GLM with diverging dimensionality, one needs to be very cautious when using p-values based on the MLE.

The key ingredients of our new geometric and probabilistic arguments are demonstrated in the proof of Theorem 2 in Section B.3. The assumption that the rescaled random design matrix $n^{-1/2}\mathbf{X}$ has the Haar measure on the Stiefel manifold $V_p(\mathbb{R}^n)$ greatly facilitates our technical analysis. The major theoretical finding is that the nonlinearity of the mean function $\boldsymbol{\mu}(\cdot)$ can be negligible in determining the asymptotic distribution of MLE as given in (11) when the dimensionality p grows at a slower rate than $n^{2/3}$, but such nonlinearity can become dominant and deform the conventional asymptotic normality when p grows at rate $n^{2/3}$ or faster. See the last paragraph of Section B.3 for more detailed in-depth discussions

1. For completeness, we present Theorem 4 in Appendix A which provides a random design version of Theorem 1 under global null and a partial converse of Theorems 2 and 3.

on such an interesting phenomenon. Furthermore, the global null assumption is a crucial component of our geometric and probabilistic argument. The global null assumption along with the distributional assumption on the design matrix ensures the symmetry property of the MLE and the useful fact that the MLE can be asymptotically independent of the random design matrix. In the absence of such an assumption, we may suspect that p-values of the noise variables can be affected by the signal variables due to asymmetry. Indeed, our simulation study in Section 4 reveals that as the number of signal variables increases, the breakdown point of the p-values occurs even earlier.

Theorem 3 further establishes that the invalidity of GLM p-values in high dimensions beyond the scenario of orthonormal design matrices considered in Theorem 2. The breakdown of the conventional p-values occurs regardless of the correlation structure of the covariates.

Our theoretical derivations detailed in the Appendix also suggest that the conventional p-values in nonlinear GLMs can generally fail to be valid when $p \sim n^{\alpha_0}$ with α_0 ranging between 1/2 and 2/3, which differs significantly from the phenomenon for linear models as discussed in the Introduction. The special feature of logistic regression model that the variance function $b''(\theta)$ takes the maximum value 1/4 at natural parameter $\theta = 0$ leads to a higher transition point of $p \sim n^{\alpha_0}$ with $\alpha_0 = 2/3$ for the case of global null $\boldsymbol{\beta}_0 = \mathbf{0}$.

4. Numerical Studies

We now investigate the breakdown point of p-values for nonlinear GLMs in diverging dimensions as predicted by our major theoretical results in Section 3 with several simulation examples. Indeed, these theoretical results are well supported by the numerical studies.

4.1. Simulation examples

Following Theorems 2–3 in Section 3, we consider three examples of the logistic regression model (1). The response vector $\mathbf{y} = (y_1, \cdots, y_n)^T$ has independent components and each y_i has Bernoulli distribution with parameter $e^{\theta_i}/(1 + e^{\theta_i})$, where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T = \mathbf{X}\boldsymbol{\beta}_0$. In example 1, we generate the $n \times p$ design matrix $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ such that $n^{-1/2}\mathbf{X}$ is uniformly distributed on the Stiefel manifold $V_p(\mathbb{R}^n)$ as in Theorem 2, while examples 2 and 3 assume that $\mathbf{X} \sim N(\mathbf{0}, I_n \otimes \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma}$ as in Theorem 3. In particular, we choose $\boldsymbol{\Sigma} = (\rho^{|j-k|})_{1 \le j,k \le p}$ with $\rho = 0, 0.5$, and 0.8 to reflect low, moderate, and high correlation levels among the covariates. Moreover, examples 1 and 2 assume the global null model with $\boldsymbol{\beta}_0 = \mathbf{0}$ following our theoretical results, whereas example 3 allows sparsity $s = \|\boldsymbol{\beta}_0\|_0$ to vary.

To examine the asymptotic results we set the sample size $n = 1000$. In each example, we consider a spectrum of dimensionality p with varying rate of growth with sample size n. As mentioned in the Introduction, the phase transition of perfect hyperplane separation happens at the point $p/n = 1/2$. Recall that Theorems 2–3 establish that the conventional GLM p-values can become invalid when $p \sim n^{2/3}$. We set $p = [n^{\alpha_0}]$ with α_0 in the grid $\{2/3 - 4\delta, \cdots, 2/3 - \delta, 2/3, 2/3 + \delta, \cdots, 2/3 + 4\delta, (\log(n) - \log(2))/\log(n)\}$ for $\delta = 0.05$. For example 3, we pick s signals uniformly at random among all but the first components, where a random half of them are chosen as 3 and the other half are set as -3.

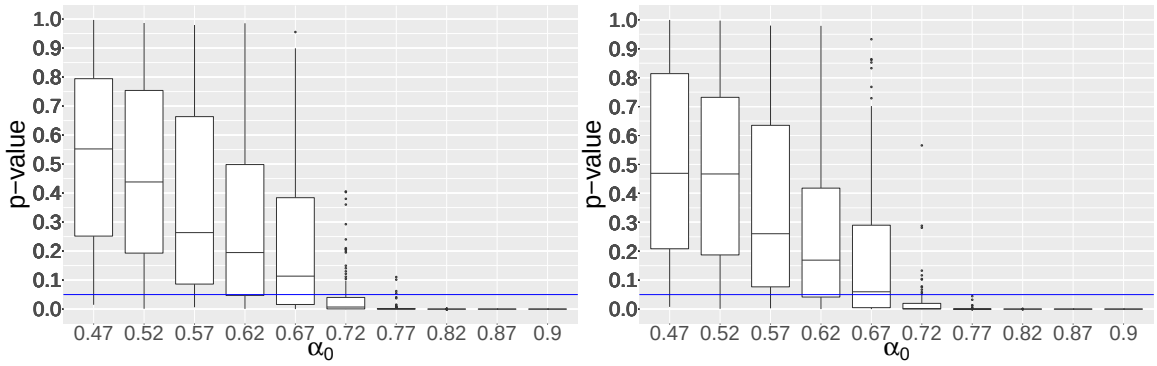
The goal of the simulation examples is to investigate empirically when the conventional GLM p-values could break down in diverging dimensions. When the asymptotic theory for

the MLE in (11) holds, the conventional p-values would be valid and distributed uniformly on the interval $[0, 1]$ under the null hypothesis. Note that the first covariate $\mathbf{x}_1$ is a null variable in each simulation example. Thus in each replication, we calculate the conventional p-value for testing the null hypothesis $H_0 : \beta_{0,1} = 0$. To check the validity of these p-values, we further test their uniformity.

For each simulation example, we first calculate the p-values for a total of $1,000$ replications as described above and then test the uniformity of these $1,000$ p-values using, for example, the Kolmogorov–Smirnov (KS) test (Kolmogorov, 1933; Smirnov, 1948) and the Anderson–Darling (AD) test (Anderson and Darling, 1952, 1954). We repeat this procedure 100 times to obtain a final set of 100 new p-values from each of these two uniformity tests. Specifically, the KS and AD test statistics for testing the uniformity on $[0, 1]$ are defined as

$$\text{KS} = \sup_{x\in[0,1]} |F_m(x) - x| \quad \text{and} \quad \text{AD} = m \int_0^1 \frac{[F_m(x) - x]^2}{x(1-x)}\, dx,$$

respectively, where $F_m(x) = m^{-1} \sum_{i=1}^m I_{(-\infty,x]}(x_i)$ is the empirical distribution function for a given sample $\{x_i\}_{i=1}^m$.

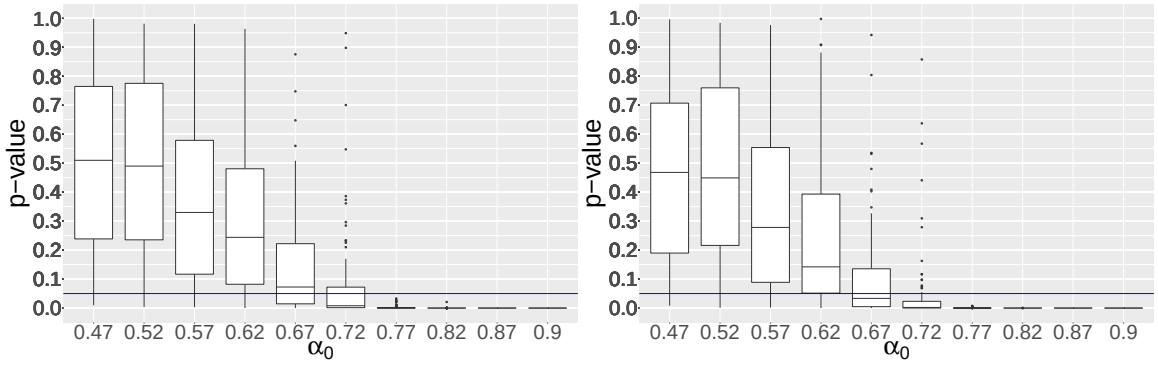


(a) KS test (b) AD test

Figure 1: Results of KS and AD tests for testing the uniformity of GLM p-values in simulation example 1 for diverging-dimensional logistic regression model with uniform orthonormal design under global null. The vertical axis represents the p-value from the KS and AD tests, and the horizontal axis stands for the growth rate α_0 of dimensionality $p = [n^{\alpha_0}]$.

4.2. Testing results

For each simulation example, we apply both KS and AD tests to verify the asymptotic theory for the MLE in (11) by testing the uniformity of conventional p-values at significance level 0.05. As mentioned in Section 4.1, we end up with two sets of 100 new p-values from the KS and AD tests. Figures 1–3 depict the boxplots of the p-values obtained from both KS and AD tests for simulation examples 1–3, respectively. In particular, we observe that the numerical results shown in Figures 1–2 for examples 1–2 are in line with our theoretical results established in Theorems 2–3, respectively, for diverging-dimensional logistic



(a) KS test for $\rho = 0.5$

(b) AD test for $\rho = 0.5$

(c) KS test for $\rho = 0.8$

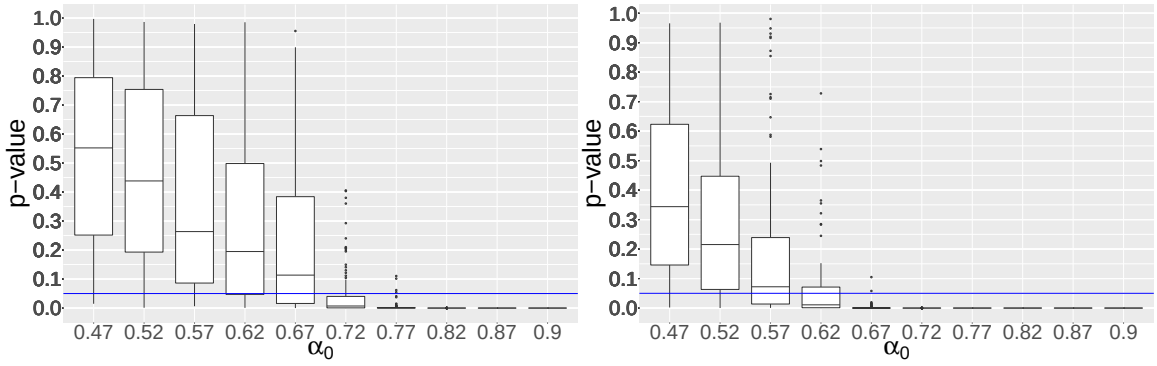
(d) AD test for $\rho = 0.8$

Figure 2: Results of KS and AD tests for testing the uniformity of GLM p-values in simulation example 2 for diverging-dimensional logistic regression model with correlated Gaussian design under global null for varying correlation level ρ. The vertical axis represents the p-value from the KS and AD tests, and the horizontal axis stands for the growth rate α_0 of dimensionality $p = [n^{\alpha_0}]$.

regression model under global null that the conventional p-values break down when $p \sim n^{\alpha_0}$ with $\alpha_0 = 2/3$. Figure 3 for example 3 examines the breakdown point of p-values with varying sparsity s. It is interesting to see that the breakdown point shifts even earlier when s increases as suggested in the discussions in Section 3.2. The results from the AD test are similar so we present only the results from the KS test for simplicity.

To gain further insights into the nonuniformity of the null p-values, we next provide an additional figure in the setting of simulation example 1. Specifically, in Figure 4 we present the histograms of the 1,000 null p-values from the first simulation repetition (out of 100) for each value of α_0. It is seen that as the dimensionality increases (i.e., α_0 increases), the null p-values have a distribution that is skewed more and more toward zero, which is prone to produce more false discoveries if these p-values are used naively in classical hypothesis testing methods.

To further demonstrate the severity of the problem, we estimate the probability of making type I error at significance level a, as the fraction of p-values below a. The means



(a) KS test for $s = 0$

(b) KS test for $s = 2$

Figure 3: Results of KS test for testing the uniformity of GLM p-values in simulation example 3 for diverging-dimensional logistic regression model with uncorrelated Gaussian design under global null for varying sparsity s. The vertical axis represents the p-value from the KS test, and the horizontal axis stands for the growth rate α_0 of dimensionality $p = [n^{\alpha_0}]$.

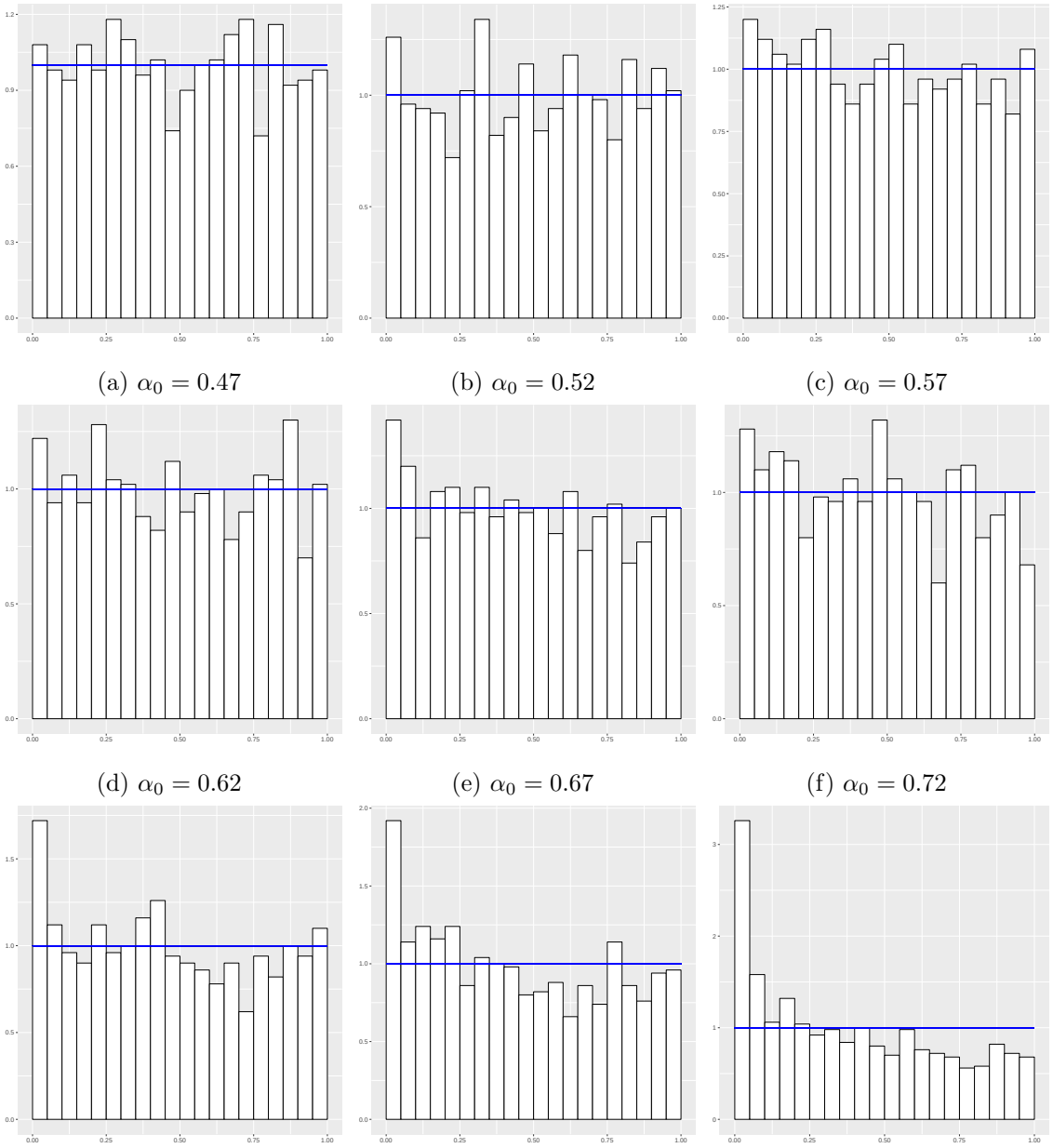
and standard deviations of the estimated probabilities are reported in Table 1 for $a = 0.05$ and 0.1. When the null p-values are distributed uniformly, the probabilities of making type I error should all be close to the target level a. However, Table 1 shows that when the growth rate of dimensionality α_0 approaches or exceeds $2/3$, these probabilities can be much larger than a, which again supports our theoretical findings. Also it is seen that when α_0 is close to but still smaller than $2/3$, the averages of estimated probabilities exceed slightly a, which could be the effect of finite sample size.

	α_0	0.10	0.47	0.57	0.67	0.77	0.87
$a = 0.05$	Mean	0.050	0.052	0.055	0.063	0.082	0.166
	SD	0.006	0.007	0.007	0.007	0.001	0.011
$a = 0.1$	Mean	0.098	0.104	0.107	0.118	0.144	0.247
	SD	0.008	0.010	0.009	0.011	0.012	0.013

Table 1: Means and standard deviations (SD) for estimated probabilities of making type I error in simulation example 1 with α_0 the growth rate of dimensionality $p = [n^{\alpha_0}]$. Two significance levels $a = 0.05$ and 0.1 are considered.

5. Discussions

In this paper we have provided characterizations of p-values in nonlinear GLMs with diverging dimensionality. The major findings are that the conventional p-values can remain valid when $p = o(n^{1/2})$, but can become invalid much earlier in nonlinear models than in linear models, where the latter case can allow for $p = o(n)$. In particular, our theoretical results



(a) $\alpha_0 = 0.47$ (b) $\alpha_0 = 0.52$ (c) $\alpha_0 = 0.57$

(d) $\alpha_0 = 0.62$ (e) $\alpha_0 = 0.67$ (f) $\alpha_0 = 0.72$

(g) $\alpha_0 = 0.77$ (h) $\alpha_0 = 0.82$ (i) $\alpha_0 = 0.87$

Figure 4: Histograms of null p-values in simulation example 1 from the first simulation repetition for different growth rates α_0 of dimensionality $p = [n^{\alpha_0}]$.

pinpoint the breakdown point of $p \sim n^{2/3}$ for p-values in diverging-dimensional logistic regression model under global null with uniform orthonormal design and correlated Gaussian design, as evidenced in the numerical results. It would be interesting to investigate such a phenomenon for more general class of random design matrices.

The problem of identifying the breakdown point of p-values becomes even more complicated and challenging when we move away from the setting of global null. Our technical analysis suggests that the breakdown point $p \sim n^{\alpha_0}$ can shift even earlier with α_0 ranging between 1/2 and 2/3. But the exact breakdown point can depend upon the number of signals s, the signal magnitude, and the correlation structure among the covariates in a rather complicated fashion. Thus more delicate mathematical analysis is needed to obtain the exact relationship. We leave such a problem for future investigation. Moving beyond the GLM setting will further complicate the theoretical analysis.

As we routinely produce p-values using algorithms, the phenomenon of nonuniformity of p-values occurring early in diverging dimensions unveiled in the paper poses useful cautions to researchers and practitioners when making decisions in real applications using results from p-value based methods. For instance, when testing the joint significance of covariates in diverging-dimensional nonlinear models, the effective sample size requirement should be checked before interpreting the testing results. Indeed, statistical inference in general high-dimensional nonlinear models is particularly challenging since obtaining accurate p-values is generally not easy. One possible route is to bypass the use of p-values in certain tasks including the false discovery rate (FDR) control; see, for example, Barber and Candès (2015); Candès et al. (2018); Fan et al. (2018) for some initial efforts made along this line.

Acknowledgments

This work was supported by NIH Grant 1R01GM131407-01, NSF CAREER Award DMS-1150318, a grant from the Simons Foundation, and Adobe Data Science Research Award. The first and last authors sincerely thank Emmanuel Candès for helpful discussions on this topic. The authors would like to thank the Associate Editor and referees for their valuable comments that helped improve the article substantially.

Appendix A. Conventional P-values in Low Dimensions under Random Design

Under the specific assumption of Gaussian design and global null $\boldsymbol{\beta}_0 = \mathbf{0}$, we can show that the asymptotic normality of MLE continues to hold without previous Conditions 1–2.

Theorem 4 *Assume that $\boldsymbol{\beta}_0 = \mathbf{0}$, the rows of $\mathbf{X}$ are i.i.d. from $N(\mathbf{0}, \boldsymbol{\Sigma})$, $b^{(5)}(\cdot)$ is uniformly bounded in its domain, and $\mathbf{y} - \boldsymbol{\mu}_0$ has uniformly sub-Gaussian components. Then if $p = O(n^{\alpha})$ with some $\alpha \in [0, 2/3)$, we have the componentwise asymptotic normality*

$$(\mathbf{A}_n^{-1})_{jj}^{-1/2} \widehat{\beta}_j \xrightarrow{\mathscr{D}} N(0, \phi),$$

where all the notation is the same as in (11).

Theorem 4 shows that the conclusions of Theorem 1 continue to hold for the case of random design and global null with the major difference that the dimensionality can be

pushed as far as $p \sim n^{2/3}$. The main reasons for presenting Theorem 4 under Gausssian design are twofold. First, Gaussian design is a widely used assumption in the literature. Second, our results on the nonuniformity of GLM p-values in diverging dimensions use geometric and probabilistic arguments which require random design setting; see Section 3 for more details. To contrast more accurately the two regimes and maintain self-contained theory, we have chosen to present Theorem 4 under Gaussian design. On the other hand, we would like to point out that Theorem 4 is not for practitioners who want to justify the usage of classical p-values. The global null assumption of $\boldsymbol{\beta}_0 = \mathbf{0}$ restricts the validity of Theorem 4 in many practical scenarios.

Appendix B. Proofs of Main Results

We provide the detailed proofs of Theorems 1–3 in this Appendix.

B.1. Proof of Theorem 1

To ease the presentation, we split the proof into two parts, where the first part locates the MLE $\widehat{\boldsymbol{\beta}}$ in an asymptotically shrinking neighborhood $\mathcal{N}_0$ of the true regression coefficient vector $\boldsymbol{\beta}_0$ with significant probability and the second part further establishes its asymptotic normality.

Part 1: Existence of a unique solution to score equation (4) in $\mathcal{N}_0$ under Condition 1 and probability bound (6). For simplicity, assume that the design matrix $\mathbf{X}$ is rescaled columnwise such that $\|\mathbf{x}_j\|_2 = \sqrt{n}$ for each $1 \le j \le p$. Consider an event

$$\mathcal{E} = \left\{ \|\boldsymbol{\xi}\|_\infty \le c_1^{-1/2} \sqrt{n \log n} \right\}, \tag{12}$$

where $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_p)^T = \mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}_0)]$. Note that for unbounded responses, the assumption of $\max_{j=1}^p \|\mathbf{x}_j\|_\infty < c_1^{1/2}\{n/(\log n)\}^{1/2}$ in Condition 1 entails that $c_1^{-1/2}\sqrt{\log n} < \min_{j=1}^p \{\|\mathbf{x}_j\|_2/\|\mathbf{x}_j\|_\infty\}$. Thus by $\|\mathbf{x}_j\|_2 = \sqrt{n}$, probability bound (6), and Bonferroni's inequality, we deduce

$$\begin{aligned} P(\mathcal{E}) &\ge 1 - \sum_{j=1}^p P\left(|\xi_j| > c_1^{-1/2} \sqrt{n \log n} \right) \\ &\ge 1 - 2pn^{-1} = 1 - O\{n^{-(1-\alpha_0)}\}, \end{aligned} \tag{13}$$

since $p = O(n^{\alpha_0})$ for some $\alpha_0 \in (0, \gamma)$ with $\gamma \in (0, 1/2]$ by assumption. Hereafter we condition on the event $\mathcal{E}$ defined in (12) which holds with significant probability.

We will show that for sufficiently large n, the score equation (4) has a solution in the neighborhood $\mathcal{N}_0$ which is a hypercube. Define two vector-valued functions

$$\boldsymbol{\gamma}(\boldsymbol{\beta}) = (\gamma_1(\boldsymbol{\beta}), \cdots, \gamma_p(\boldsymbol{\beta}))^T = \mathbf{X}^T \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})$$

and

$$\boldsymbol{\Psi}(\boldsymbol{\beta}) = \boldsymbol{\gamma}(\boldsymbol{\beta}) - \boldsymbol{\gamma}(\boldsymbol{\beta}_0) - \boldsymbol{\xi}, \quad \boldsymbol{\beta} \in \mathbb{R}^p.$$

Then equation (4) is equivalent to $\boldsymbol{\Psi}(\boldsymbol{\beta}) = \mathbf{0}$. We need to show that the latter has a solution inside the hypercube $\mathcal{N}_0$. To this end, applying a second order Taylor expansion of $\boldsymbol{\gamma}(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_0$ with the Lagrange remainder term componentwise leads to

$$\boldsymbol{\gamma}(\boldsymbol{\beta}) = \boldsymbol{\gamma}(\boldsymbol{\beta}_0) + \mathbf{X}^T \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right) \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \mathbf{r}, \tag{14}$$

where $\mathbf{r} = (r_1, \cdots, r_p)^T$ and for each $1 \le j \le p$,

$$r_j = \frac{1}{2} \left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)^T \nabla^2 \gamma_j(\boldsymbol{\beta}_j) \left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)$$

with $\boldsymbol{\beta}_j$ some p-dimensional vector lying on the line segment joining $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. It follows from (9) in Condition 1 that

$$\begin{aligned} \|\mathbf{r}\|_\infty &\le \max_{\boldsymbol{\delta} \in \mathcal{N}_0} \max_{j=1}^{p} \frac{1}{2} \lambda_{\max} \left[\mathbf{X}^T \mathrm{diag}\left\{ |\mathbf{x}_j| \circ \left|\boldsymbol{\mu}''\left(\mathbf{X}\boldsymbol{\delta}\right)\right| \right\} \mathbf{X}\right] \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 \\ &= O\left\{ p n^{1-2\gamma} (\log n)^2 \right\}. \end{aligned} \tag{15}$$

Let us define another vector-valued function

$$\overline{\boldsymbol{\Psi}}(\boldsymbol{\beta}) \equiv \left[\mathbf{X}^T \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right) \mathbf{X}\right]^{-1} \boldsymbol{\Psi}(\boldsymbol{\beta}) = \boldsymbol{\beta} - \boldsymbol{\beta}_0 + \mathbf{u}, \tag{16}$$

where $\mathbf{u} = -[\mathbf{X}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{X}]^{-1}(\boldsymbol{\xi} - \mathbf{r})$. It follows from (12), (15), and (8) in Condition 1 that for any $\boldsymbol{\beta} \in \mathcal{N}_0$,

$$\begin{aligned} \|\mathbf{u}\|_\infty &\le \left\| \left[\mathbf{X}^T \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right) \mathbf{X}\right]^{-1} \right\|_\infty \left(\|\boldsymbol{\xi}\|_\infty + \|\mathbf{r}\|_\infty\right) \\ &= O\left[b_n n^{-1/2} \sqrt{\log n} + b_n p n^{-2\gamma} (\log n)^2 \right]. \end{aligned} \tag{17}$$

By the assumptions of $p = O(n^{\alpha_0})$ with constant $\alpha_0 \in (0, \gamma)$ and $b_n = o\{\min(n^{1/2-\gamma}\sqrt{\log n}, n^{2\gamma - \alpha_0 - 1/2}/(\log n)^2\}$, we have

$$\|\mathbf{u}\|_\infty = o(n^{-\gamma} \log n).$$

Thus in light of (16), it holds for large enough n that when $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)_j = n^{-\gamma}\sqrt{\log n}$,

$$\overline{\Psi}_j(\boldsymbol{\beta}) \ge n^{-\gamma}\sqrt{\log n} - \|\mathbf{u}\|_\infty \ge 0, \tag{18}$$

and when $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)_j = -n^{-\gamma}\sqrt{\log n}$,

$$\overline{\Psi}_j(\boldsymbol{\beta}) \le -n^{-\gamma}\sqrt{\log n} + \|\mathbf{u}\|_\infty \le 0, \tag{19}$$

where $\overline{\boldsymbol{\Psi}}(\boldsymbol{\beta}) = (\overline{\Psi}_1(\boldsymbol{\beta}), \cdots, \overline{\Psi}_p(\boldsymbol{\beta}))^T$.

By the continuity of the vector-valued function $\overline{\boldsymbol{\Psi}}(\boldsymbol{\beta})$, (18), and (19), Miranda's existence theorem Vrahatis (1989) ensures that equation $\overline{\boldsymbol{\Psi}}(\boldsymbol{\beta}) = \mathbf{0}$ has a solution $\widehat{\boldsymbol{\beta}}$ in $\mathcal{N}_0$. Clearly, $\widehat{\boldsymbol{\beta}}$ also solves equation $\boldsymbol{\Psi}(\boldsymbol{\beta}) = \mathbf{0}$ in view of (16). Therefore, we have shown that score equation (4) indeed has a solution $\widehat{\boldsymbol{\beta}}$ in $\mathcal{N}_0$. The strict concavity of the log-likelihood function (2) by assumptions for model (1) entails that $\widehat{\boldsymbol{\beta}}$ is the MLE.

Part 2: Conventional asymptotic normality of the MLE $\widehat{\boldsymbol{\beta}}$. Fix any $1 \le j \le p$. In light of (16), we have $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \mathbf{A}_n^{-1}(\boldsymbol{\xi} - \mathbf{r})$, which results in

$$(\mathbf{A}_n^{-1})_{jj}^{-1/2}(\widehat{\beta}_j - \beta_{0,j}) = (\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1}\boldsymbol{\xi} - (\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1}\mathbf{r} \tag{20}$$

with $\mathbf{e}_j \in \mathbb{R}^p$ having one for the jth component and zero otherwise. Note that since the smallest and largest eigenvalues of $n^{-1}\mathbf{A}_n$ are bounded away from 0 and ∞ by Condition 2, it is easy to show that $(\mathbf{A}_n^{-1})_{jj}^{-1/2}$ is of exact order $n^{1/2}$. In view of (17), it holds on the event $\mathcal{E}$ defined in (12) that

$$\begin{aligned} \left\|\mathbf{A}_n^{-1}\mathbf{r}\right\|_\infty &\le \left\|\left[\mathbf{X}^T\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\mathbf{X}\right]^{-1}\right\|_\infty \|\mathbf{r}\|_\infty \\ &= O\left[b_n p n^{-2\gamma}(\log n)^2\right] = o(n^{-1/2}), \end{aligned}$$

since $b_n = o\{n^{2\gamma-\alpha_0-1/2}/(\log n)^2\}$ by assumption. This leads to

$$(\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1}\mathbf{r} = O(n^{1/2}) \cdot o_P(n^{-1/2}) = o_P(1). \tag{21}$$

It remains to consider the term $(\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1}\boldsymbol{\xi} = \sum_{i=1}^n \eta_i$, where $\eta_i = (\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T \mathbf{A}_n^{-1}\mathbf{z}_i[y_i - b'(\theta_{0,i})]$. Clearly, the n random variables η_i's are independent with mean 0 and

$$\sum_{i=1}^n \mathrm{var}(\eta_i) = (\mathbf{A}_n^{-1})_{jj}^{-1}\mathbf{e}_j^T\mathbf{A}_n^{-1}(\phi\mathbf{A}_n)\mathbf{A}_n^{-1}\mathbf{e}_j = \phi.$$

It follows from Condition 2 and the Cauchy–Schwarz inequality that

$$\begin{aligned} \sum_{i=1}^n E\left|\eta_i\right|^3 &= \sum_{i=1}^n \left|(\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1}\mathbf{z}_i\right|^3 E\left|y_i - b'\left(\theta_{0,i}\right)\right|^3 \\ &= O(1)\sum_{i=1}^n \left|(\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1}\mathbf{z}_i\right|^3 \\ &\le O(1)\sum_{i=1}^n \left\|(\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1/2}\right\|_2^3 \left\|\mathbf{A}_n^{-1/2}\mathbf{z}_i\right\|_2^3 \\ &= O(1)\sum_{i=1}^n \left(\mathbf{z}_i^T\mathbf{A}_n^{-1}\mathbf{z}_i\right)^{3/2} = o(1). \end{aligned}$$

Thus an application of Lyapunov's theorem yields

$$(\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1}\boldsymbol{\xi} = \sum_{i=1}^n \eta_i \xrightarrow{\mathscr{D}} N(0,\phi). \tag{22}$$

By Slutsky's lemma, we see from (20)–(22) that

$$(\mathbf{A}_n^{-1})_{jj}^{-1/2}(\widehat{\beta}_j - \beta_{0,j}) \xrightarrow{\mathscr{D}} N(0,\phi),$$

showing the asymptotic normality of each component $\widehat{\beta}_j$ of the MLE $\widehat{\boldsymbol{\beta}}$.

We further establish the asymptotic normality for the one-dimensional projections of the MLE $\widehat{\boldsymbol{\beta}}$. Fix an arbitrary vector $\mathbf{u} \in \mathbb{R}^p$ with $\|\mathbf{u}\|_2 = 1$ satisfying the L_1 sparsity bound $\|\mathbf{u}\|_1 = O(s_n)$. In light of (16), we have $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \mathbf{A}_n^{-1}(\boldsymbol{\xi} - \mathbf{r})$, which results in

$$(\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{u})^{-1/2}(\mathbf{u}^T\widehat{\boldsymbol{\beta}} - \mathbf{u}^T\boldsymbol{\beta}_0) = (\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{u})^{-1/2}\mathbf{u}^T\mathbf{A}_n^{-1}\boldsymbol{\xi} - (\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{u})^{-1/2}\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{r}. \quad (23)$$

Note that since the smallest and largest eigenvalues of $n^{-1}\mathbf{A}_n$ are bounded away from 0 and ∞ by Condition 2, it is easy to show that $(\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{u})^{-1/2}$ is of exact order $n^{1/2}$. In view of (17), it holds on the event $\mathcal{E}$ defined in (12) that

$$\begin{aligned} \left\|\mathbf{A}_n^{-1}\mathbf{r}\right\|_\infty &\leq \left\|\left[\mathbf{X}^T\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\mathbf{X}\right]^{-1}\right\|_\infty \|\mathbf{r}\|_\infty \\ &= O\left\{b_n p n^{-2\gamma}(\log n)^2\right\} = o(s_n^{-1}n^{-1/2}) \end{aligned}$$

since $b_n = o\{s_n^{-1}n^{2\gamma-\alpha_0-1/2}/(\log n)^2\}$ by assumption. This leads to

$$(\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{u})^{-1/2}\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{r} = O(n^{1/2}) \cdot \|\mathbf{u}\|_1 \cdot \|\mathbf{A}_n^{-1}\mathbf{r}\|_\infty = o_P(1) \quad (24)$$

since $\|\mathbf{u}\|_1 = O(s_n)$ by assumption.

It remains to consider the term $(\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{u})^{-1/2}\mathbf{u}^T\mathbf{A}_n^{-1}\boldsymbol{\xi} = \sum_{i=1}^n \eta_i$ with $\eta_i = (\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{u})^{-1/2}$ $\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{z}_i[y_i - b'(\theta_{0,i})]$. Clearly, the n random variables η_i's are independent with mean 0 and

$$\sum_{i=1}^n \mathrm{var}(\eta_i) = (\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{u})^{-1}\mathbf{u}^T\mathbf{A}_n^{-1}(\phi\mathbf{A}_n)\mathbf{A}_n^{-1}\mathbf{u} = \phi.$$

It follows from Condition 2 and the Cauchy–Schwarz inequality that

$$\begin{aligned} \sum_{i=1}^n E\left|\eta_i\right|^3 &= \sum_{i=1}^n \left|(\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{u})^{-1/2}\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{z}_i\right|^3 E\left|y_i - b'\left(\theta_{0,i}\right)\right|^3 \\ &= O(1)\sum_{i=1}^n \left|(\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{u})^{-1/2}\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{z}_i\right|^3 \\ &\leq O(1)\sum_{i=1}^n \left\|(\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{u})^{-1/2}\mathbf{u}^T\mathbf{A}_n^{-1/2}\right\|_2^3 \left\|\mathbf{A}_n^{-1/2}\mathbf{z}_i\right\|_2^3 \\ &= O(1)\sum_{i=1}^n \left(\mathbf{z}_i^T\mathbf{A}_n^{-1}\mathbf{z}_i\right)^{3/2} = o(1). \end{aligned}$$

Thus an application of Lyapunov's theorem yields

$$(\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{u})^{-1/2}\mathbf{u}^T\mathbf{A}_n^{-1}\boldsymbol{\xi} = \sum_{i=1}^n \eta_i \xrightarrow{\mathscr{D}} N(0,\phi). \quad (25)$$

By Slutsky's lemma, we see from (23)–(25) that

$$(\mathbf{u}^T\mathbf{A}_n^{-1}\mathbf{u})^{-1/2}(\mathbf{u}^T\widehat{\beta} - \mathbf{u}^T\boldsymbol{\beta}_0) \xrightarrow{\mathscr{D}} N(0,\phi),$$

showing the asymptotic normality of any L_1-sparse one-dimensional projection $\mathbf{u}^T\widehat{\boldsymbol{\beta}}$ of the MLE $\widehat{\boldsymbol{\beta}}$. This completes the proof of Theorem 1.

B.2. Proof of Theorem 4

The proof is similar to that for Theorem 1. Without loss of generality, we assume that $\mathbf{\Sigma} = I_p$ because under global null, a rotation of $\mathbf{X}$ yields standard normal rows. First let $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_p)^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}_0]$, where $\boldsymbol{\mu}_0 = b'(0)\mathbf{1}$ with $\mathbf{1} = (1, \cdots, 1)^T \in \mathbb{R}^n$ because $\boldsymbol{\beta}_0 = 0$. Then $\mathbf{y} - \boldsymbol{\mu}_0$ has i.i.d. uniform sub-Gaussian components and is independent of $\mathbf{X} = (\mathbf{z}_1, \cdots, \mathbf{z}_p) \in \mathbb{R}^{n\times p}$. Define event

$$\mathcal{E} = \{\|\boldsymbol{\xi}\|_\infty \le c_2\sqrt{n^{-1}\log n}\}.$$

By Lemma 5, it is seen that $P(\mathcal{E}) \ge 1 - o(p^{-a})$. Furthermore, define the neighborhood

$$\mathcal{N}_0 = \{\|\boldsymbol{\beta}\|_\infty \le c_3\sqrt{n^{-1}\log n}\} \tag{26}$$

for some $c_3 > c_2(b''(0))^{-1}$. We next show that the MLE must fall into the region $\mathcal{N}_0$ with probability at least $1 - O(p^{-a})$ following the similar arguments in Theorem 1.

First, we define

$$\boldsymbol{\gamma}(\boldsymbol{\beta}) = (\gamma_1(\boldsymbol{\beta}), \cdots, \gamma_p(\boldsymbol{\beta}))^T \equiv \mathbf{X}^T\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})$$

and

$$\mathbf{\Psi}(\boldsymbol{\beta}) = \boldsymbol{\gamma}(\boldsymbol{\beta}) - \boldsymbol{\gamma}(\boldsymbol{\beta}_0) - \mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}_0], \quad \boldsymbol{\beta} \in \mathbb{R}^p.$$

Applying a forth order Taylor expansion of $\boldsymbol{\gamma}(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_0 = \mathbf{0}$ with the Lagrange remainder term componentwise leads to

$$\boldsymbol{\gamma}(\boldsymbol{\beta}) = \boldsymbol{\gamma}(\boldsymbol{\beta}_0) + b''(0)\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \mathbf{r} + \mathbf{s} + \mathbf{t},$$

where $\mathbf{r} = (r_1, \cdots, r_p)^T$, $\mathbf{s} = (s_1, \cdots, s_p)^T$, $\mathbf{t} = (t_1, \cdots, t_p)^T$ and for each $1 \le j \le p$,

$$r_j = \frac{b'''(0)}{2}\sum_{i=1}^n x_{ij}(\mathbf{x}_i^T\boldsymbol{\beta})^2 \tag{27}$$

$$s_j = \frac{b^{(4)}(0)}{6}\sum_{i=1}^n x_{ij}(\mathbf{x}_i^T\boldsymbol{\beta})^3 \tag{28}$$

$$t_j = \frac{1}{24}\sum_{i=1}^n b^{(5)}(\mathbf{x}_i^T\widetilde{\boldsymbol{\beta}^j})x_{ij}(\mathbf{x}_i^T\boldsymbol{\beta})^4. \tag{29}$$

with $\widetilde{\boldsymbol{\beta}^j}$ some p-dimensional vector lying on the line segment joining $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$.

Let us define another vector-valued function

$$\overline{\mathbf{\Psi}}(\boldsymbol{\beta}) \equiv \left[b''(0)\mathbf{X}^T\mathbf{X}\right]^{-1}\mathbf{\Psi}(\boldsymbol{\beta}) = \boldsymbol{\beta} - \boldsymbol{\beta}_0 + \mathbf{u}, \tag{30}$$

where $\mathbf{u} = -(b''(0))^{-1}\boldsymbol{\xi} + [b''(0)\mathbf{X}^T\mathbf{X}]^{-1}(\mathbf{r} + \mathbf{s} + \mathbf{t})$. It follows from the above derivation that for any $\boldsymbol{\beta} \in \mathcal{N}_0$,

$$\|\mathbf{u}\|_\infty \le \left\|(b''(0))^{-1}\boldsymbol{\xi}\right\|_\infty + \left\|\left[b''(0)\mathbf{X}^T\mathbf{X}\right]^{-1}(\mathbf{r} + \mathbf{s} + \mathbf{t})\right\|_\infty.$$

Now, we bound the terms on the right hand side.

First note that on event $\mathcal{E}$,

$$\left\|(b''(0))^{-1}\boldsymbol{\xi}\right\|_\infty \le (b''(0))^{-1}c_2\sqrt{n\log n}. \tag{31}$$

Then, we consider the next term: $\left\|\left[b''(0)\mathbf{X}^T\mathbf{X}\right]^{-1}(\mathbf{r}+\mathbf{s}+\mathbf{t})\right\|_\infty$. We observe that

$$\begin{aligned}\left\|\left[b''(0)\mathbf{X}^T\mathbf{X}\right]^{-1}(\mathbf{r}+\mathbf{s}+\mathbf{t})\right\|_\infty &\le |b''(0)|^{-1}\left\|\left[n^{-1}\mathbf{X}^T\mathbf{X}\right]^{-1}\right\|_\infty \left\|n^{-1}(\mathbf{r}+\mathbf{s}+\mathbf{t})\right\|_\infty \\ &\le |b''(0)|^{-1}\left\|\left[n^{-1}\mathbf{X}^T\mathbf{X}\right]^{-1}\right\|_\infty \\ &\quad\cdot\left(\left\|n^{-1}\mathbf{r}\right\|_\infty + \left\|n^{-1}\mathbf{s}\right\|_\infty + \left\|n^{-1}\mathbf{t}\right\|_\infty\right).\end{aligned}$$

By Lemma 6, we have that $\left\|\left[n^{-1}\mathbf{X}^T\mathbf{X}\right]^{-1}\right\|_\infty \le 1 + O(pn^{-1/2})$. Lemmas 10, 11, 12 assert that

$$\begin{aligned}&\left(\left\|n^{-1}\mathbf{r}\right\|_\infty + \left\|n^{-1}\mathbf{s}\right\|_\infty + \left\|n^{-1}\mathbf{t}\right\|_\infty\right) \\ &\quad= \{n^{\alpha-5/6}\log n + n^{3/2\alpha-5/4}(\log n)^{3/2} + n^{\alpha-1}(\log n)^{1/2} + n^{2\alpha-3/2}(\log n)^{3/2}\}\sqrt{n^{-1}\log n}.\end{aligned}$$

We combine last two bounds so that we have

$$\left\|\left[b''(0)\mathbf{X}^T\mathbf{X}\right]^{-1}(\mathbf{r}+\mathbf{s}+\mathbf{t})\right\|_\infty = o(\sqrt{n^{-1}\log n}) \tag{32}$$

with probability at least $1-o(p^{-c})$ when $p = O(n^\alpha)$ with $\alpha < 2/3$.

Combining equations (31) and (32), we obtain that if $p = O(n^\alpha)$ with $\alpha \in [0, 2/3)$, then

$$\|\mathbf{u}\|_\infty \le c_3\sqrt{n^{-1}\log n}.$$

Thus, the MLE must fall into the region $\mathcal{N}_0$ following the similar arguments in Theorem 1.

Next, we show the componentwise asymptotic normality of the MLE $\widehat{\boldsymbol{\beta}}$. By equation (30), we have $\widehat{\boldsymbol{\beta}} = -\mathbf{u} = (b''(0))^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T[\mathbf{y}-\boldsymbol{\mu}_0] - [b''(0)\mathbf{X}^T\mathbf{X}]^{-1}(\mathbf{r}+\mathbf{s}+\mathbf{t})$. So, we can write

$$\hat{\beta}_j = (b''(0))^{-1}n^{-1}\mathbf{e}_j^T\mathbf{X}^T[\mathbf{y}-\boldsymbol{\mu}_0] + (b''(0))^{-1}T - \mathbf{e}_j^T[b''(0)\mathbf{X}^T\mathbf{X}]^{-1}(\mathbf{r}+\mathbf{s}+\mathbf{t}) \tag{33}$$

where $T = \mathbf{e}_j^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T[\mathbf{y}-\boldsymbol{\mu}_0] - n^{-1}\mathbf{e}_j^T\mathbf{X}^T[\mathbf{y}-\boldsymbol{\mu}_0]$. By Lemma 13 and Equation (32), both $n^{1/2}(b''(0))^{-1}T$ and $n^{1/2}\mathbf{e}_j^T[b''(0)\mathbf{X}^T\mathbf{X}]^{-1}(\mathbf{r}+\mathbf{s}+\mathbf{t})$ converges to zero in probability. So, it is enough to consider the first summand in (33). Now, we show that $n^{-1/2}\mathbf{e}_j^T\mathbf{X}^T[\mathbf{y}-\boldsymbol{\mu}_0]$ is asymptotically normal. In fact, we can write $\mathbf{e}_j^T\mathbf{X}^T[\mathbf{y}-\boldsymbol{\mu}_0] = \sum_{i=1}^n x_{ij}y_i$ where each summand $x_{ij}y_i$ is independent over i and has variance $\phi b''(0)$. Moreover, $\sum_{i=1}^n E|x_{ij}y_i|^3 = O(n)$ since $|x_{ij}|^3$ and $|y_i|^3$ are independent and finite mean. So, we apply Lyapunov's theorem to obtain $b''(0)^{-1/2}n^{-1/2}\mathbf{e}_j^T\mathbf{X}^T[\mathbf{y}-\boldsymbol{\mu}_0] \xrightarrow{\mathscr{D}} N(0,\phi)$. Finally, we know that $b''(0)n(\mathbf{A}_n^{-1})_{jj} \to 1$ in probability from the remark in Theorem 1. Thus, Slutsky's lemma yields

$$(\mathbf{A}_n^{-1})_{jj}^{-1/2}\hat{\beta}_j \xrightarrow{\mathscr{D}} N(0,\phi). \tag{34}$$

This completes the proof of the theorem.

Lemma 5 *Assume that the components of* $\mathbf{y}-\boldsymbol{\mu}_0$ *are uniform sub-Gaussians. That is, there exist a positive constant* C *such that* $P(|(\mathbf{y}-\boldsymbol{\mu}_0)_i| > t) \le C\exp\{-Ct^2\}$ *for all* $1 \le i \le n$. *Then, it holds that, for some positive constant* c_2,

$$\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}_0)\|_\infty \le c_2\sqrt{n^{-1}\log n}.$$

with asymptotic probability $1-o(p^{-a})$.

Proof We prove the result by conditioning on $\mathbf{X}$. Let $\mathbf{E} = n^{-1}\mathbf{X}^T\mathbf{X} - \mathbf{I}_p$. Then by matrix inversion,

$$\begin{aligned}(n^{-1}\mathbf{X}^T\mathbf{X})^{-1} &= (\mathbf{I}_p+\mathbf{E})^{-1} = \mathbf{I}_p - \sum_{k=1}^{\infty}(-1)^{k+1}(\mathbf{E})^k \\ &= \mathbf{I}_p - \mathbf{E} + \sum_{k=2}^{\infty}(-1)^k(\mathbf{E})^k = 2\mathbf{I}_p - n^{-1}\mathbf{X}^T\mathbf{X} + \sum_{k=2}^{\infty}(-1)^k(\mathbf{E})^k.\end{aligned}$$

Thus, it follows that

$$\begin{aligned}&\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}_0)\|_\infty \\ &\le \|2n^{-1}\mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}_0)\|_\infty + \|n^{-2}\mathbf{X}^T\mathbf{X}\mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}_0)\|_\infty + \left\|n^{-1}\sum_{k=2}^{\infty}(-1)^k(\mathbf{E})^k\mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}_0)\right\|_\infty \\ &= \eta_1+\eta_2+\eta_3.\end{aligned}$$

In the rest of the proof, we will bound η_1, η_2 and η_3.

Part 1: Bound of η_1.

First, it is easy to see that

$$\begin{aligned}\eta_1 &= \|2n^{-1}\mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}_0)\|_\infty \\ &= 2\max_{1\le j\le p}\left|n^{-1}\sum_{i=1}^{n}x_{ij}(\mathbf{y}-\boldsymbol{\mu}_0)_i\right|.\end{aligned}$$

We observe that each summand $x_{ij}(\mathbf{y}-\boldsymbol{\mu}_0)_i$ is the product of two subgaussian random variables, and so satisfies $P(|x_{ij}(\mathbf{y}-\boldsymbol{\mu}_0)_i| > t) \le C\exp(-Ct)$ for some constant $C > 0$ by Lemma 1 in Fan et al. (2016). Moreover, $E[x_{ij}(\mathbf{y}-\boldsymbol{\mu}_0)_i] = 0$ since x_{ij} and $(\mathbf{y}-\boldsymbol{\mu}_0)_i$ are independent and have zero mean. Thus, we can use Lemma 9 by setting $W_{ij} = x_{ij}(\mathbf{y}-\boldsymbol{\mu}_0)_i$ and $\alpha = 1$. So, we get

$$\eta_1 = 2\max_{1\le j\le p}\left|n^{-1}\sum_{i=1}^{n}x_{ij}(\mathbf{y}-\boldsymbol{\mu}_0)_i\right| \le c_2\sqrt{n^{-1}\log p} \tag{35}$$

with probability $1-O(p^{-c})$ for some positive constants c and c_2.

Part 2: Bound of η_2.

Now, we study $\eta_2 = \|n^{-2}\mathbf{X}^T\mathbf{X}\mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}_0)\|_\infty$. Let $\mathbf{z}_k$ be the k-th column of X, that is $\mathbf{z}_k = \mathbf{X}\mathbf{e}_k$. Direct calculations yield

$$\mathbf{e}_k^T\mathbf{X}^T\mathbf{X}\mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}_0) = \sum_{j=1}^{p}(\mathbf{z}_k^T\mathbf{z}_j)(\mathbf{z}_j^T(\mathbf{y}-\boldsymbol{\mu}_0)) = \|\mathbf{z}_k\|_2^2\mathbf{z}_k^T(\mathbf{y}-\boldsymbol{\mu}_0) + \sum_{j\ne k}^{p}(\mathbf{z}_k^T\mathbf{z}_j)(\mathbf{z}_j^T(\mathbf{y}-\boldsymbol{\mu}_0)).$$

Thus, it follows that

$$\left\|(\mathbf{X}^T\mathbf{X}\mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}_0)\right\|_\infty \le \max_k \left|\|\mathbf{z}_k\|_2^2\mathbf{z}_k^T(\mathbf{y}-\boldsymbol{\mu}_0)\right| + \max_k \left|\sum_{j\neq k}^{p}(\mathbf{z}_k^T\mathbf{z}_j)(\mathbf{z}_j^T(\mathbf{y}-\boldsymbol{\mu}_0))\right|. \quad (36)$$

First, we consier $\max_k \left|\|\mathbf{z}_k\|_2^2\mathbf{z}_k^T(\mathbf{y}-\boldsymbol{\mu}_0)\right|$. Lemma 14 shows that $\max_k \|\mathbf{z}_k\|_2^2 \le O(n)$ with probability $1-O(p^{-c})$. We also have $\max_k \left|\mathbf{z}_k^T(\mathbf{y}-\boldsymbol{\mu}_0)\right| = \frac{n}{2}\eta_1 \le O(\sqrt{n\log p})$ by equation (35). It follows that

$$\max_k \left|\|\mathbf{z}_k\|_2^2\mathbf{z}_k^T(\mathbf{y}-\boldsymbol{\mu}_0)\right| \le \max_k \|\mathbf{z}_k\|_2^2 \max_k \left|\mathbf{z}_k^T(\mathbf{y}-\boldsymbol{\mu}_0)\right| \le O(n\sqrt{n\log p}). \quad (37)$$

Next, let $a_j = \mathbf{z}_k^T\mathbf{z}_j/\|\mathbf{z}_k\|_2$ and $b_j = \mathbf{z}_j^T(\mathbf{y}-\boldsymbol{\mu}_0)/\|\mathbf{y}-\boldsymbol{\mu}_0\|_2$. Then it is easy to see that conditional on $\mathbf{z}_k$ and $\mathbf{y}$, $a_j \sim N(0,1)$, $b_j \sim N(0,1)$ and $\mathrm{cov}(a_j, b_j|\mathbf{z}_k,\mathbf{y}) = \mathbf{z}_k^T(\mathbf{y}-\boldsymbol{\mu}_0)/(\|\mathbf{z}_k\|_2\|\mathbf{y}-\boldsymbol{\mu}_0\|_2)$. By (E.6) of Lemma 7 in Fan et al. (2016), it can be shown that

$$\begin{aligned}
&P\left(\frac{1}{p-1}\left|\sum_{j\neq k}^{p}(\mathbf{z}_k^T\mathbf{z}_j)(\mathbf{z}_j^T(\mathbf{y}-\boldsymbol{\mu}_0)) - \mathbf{z}_k^T(\mathbf{y}-\boldsymbol{\mu}_0)\right| \ge c\|\mathbf{z}_k\|_2\|\mathbf{y}-\boldsymbol{\mu}_0\|_2\sqrt{p^{-1}\log p}\Big|\mathbf{z}_k,\mathbf{y}\right)\\
&= P\left(\frac{1}{p-1}\left|\sum_{j\neq k}^{p} a_jb_j - \frac{\mathbf{z}_k^T(\mathbf{y}-\boldsymbol{\mu}_0)}{\|\mathbf{z}_k\|_2\|\mathbf{y}-\boldsymbol{\mu}_0\|_2}\right| \ge c\sqrt{p^{-1}\log p}\Big|\mathbf{z}_k,\mathbf{y}\right) \le cp^{-c_1},
\end{aligned}$$

where c_1 is some large positive constant independent of $\mathbf{z}_k$ and $\mathbf{y}$. Moreover, we can choose c_1 as large as we want by increasing c. Thus, it follows that

$$P\left(\frac{1}{p-1}\left|\sum_{j\neq k}^{p}(\mathbf{z}_k^T\mathbf{z}_j)(\mathbf{z}_j^T(\mathbf{y}-\boldsymbol{\mu}_0)) - \mathbf{z}_k^T(\mathbf{y}-\boldsymbol{\mu}_0)\right| \ge c\|\mathbf{z}_k\|_2\|\mathbf{y}-\boldsymbol{\mu}_0\|_2\sqrt{p^{-1}\log p}\right) \le cp^{-c_1}.$$

It follows from probability union bound that

$$P\left(\frac{1}{p-1}\max_k\left|\frac{1}{\|\mathbf{z}_k\|_2\|\mathbf{y}-\boldsymbol{\mu}_0\|_2}\sum_{j\neq k}^{p}(\mathbf{z}_k^T\mathbf{z}_j)(\mathbf{z}_j^T(\mathbf{y}-\boldsymbol{\mu}_0)) - \mathbf{z}_k^T(\mathbf{y}-\boldsymbol{\mu}_0)\right| \ge c\sqrt{p^{-1}\log p}\right) \le cp^{-c_1+1}.$$

Taking $c_1 > 1$ yields that with probability at least $1-o(p^{-a})$ for some $a>0$,

$$\max_k\left\{\frac{1}{\|\mathbf{z}_k\|_2\|\mathbf{y}-\boldsymbol{\mu}_0\|_2}\left|\frac{1}{p-1}\sum_{j\neq k}^{p}(\mathbf{z}_k^T\mathbf{z}_j)(\mathbf{z}_j^T(\mathbf{y}-\boldsymbol{\mu}_0)) - \mathbf{z}_k^T(\mathbf{y}-\boldsymbol{\mu}_0)\right|\right\} \le c\sqrt{p^{-1}\log p}.$$

By Lemma 14, we have $\max_k \|\mathbf{z}_k\|_2 = \sqrt{\max_k \|\mathbf{z}_k\|_2^2} \leq O_p(\sqrt{n})$. Therefore, by using the fact that $\|\mathbf{y} - \boldsymbol{\mu}_0\|_2 \leq O_p(\sqrt{n})$, we have

$$\begin{aligned}
&\max_k \left| \sum_{j \neq k}^{p} (\mathbf{z}_k^T \mathbf{z}_j)(\mathbf{z}_j^T(\mathbf{y} - \boldsymbol{\mu}_0)) \right| \\
&\leq \max_k \left| \sum_{j \neq k}^{p} [(\mathbf{z}_k^T \mathbf{z}_j)(\mathbf{z}_j^T(\mathbf{y} - \boldsymbol{\mu}_0))] - (p-1)\mathbf{z}_k^T(\mathbf{y} - \boldsymbol{\mu}_0) \right| + (p-1) \max_k |\mathbf{z}_k^T(\mathbf{y} - \boldsymbol{\mu}_0)| \\
&\leq p \max_k \|\mathbf{z}_k\|_2 \|\mathbf{y} - \boldsymbol{\mu}_0\|_2 \max_k \left\{ \frac{1}{\|\mathbf{z}_k\|_2 \|\mathbf{y} - \boldsymbol{\mu}_0\|_2} \left| \frac{1}{p-1} \sum_{j \neq k}^{p} [(\mathbf{z}_k^T \mathbf{z}_j)(\mathbf{z}_j^T(\mathbf{y} - \boldsymbol{\mu}_0)) - \mathbf{z}_k^T(\mathbf{y} - \boldsymbol{\mu}_0)] \right| \right\} \\
&\quad + p \max_k |\mathbf{z}_k^T(\mathbf{y} - \boldsymbol{\mu}_0)| \\
&\leq cpn\sqrt{\log p}\sqrt{p^{-1}\log p} + cp\sqrt{n \log p}. \qquad (38)
\end{aligned}$$

Combining (36)–(38) yields

$$\eta_2 = \|n^{-2}\mathbf{X}^T\mathbf{X}\mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu}_0)\|_\infty \leq cp^{1/2}n^{-1}\log p = o(\sqrt{n^{-1}\log n}). \qquad (39)$$

Part 3: Bound of η_3.

Finally, we study η_3. We observe that $\eta_3 \leq \left\|\sum_{k=2}^{\infty}(-1)^{k+1}(\mathbf{E})^k\right\|_\infty \|n^{-1}\mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu}_0)\|_\infty$. Lemma 7 proves that $\left\|\sum_{k=2}^{\infty}(-1)^{k+1}(\mathbf{E})^k\right\|_\infty \leq O(p^{3/2}n^{-1})$ while equation (35) shows that $\|n^{-1}\mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu}_0)\|_\infty = O(\sqrt{n^{-1}\log p})$ with probability $1 - O(p^{-c})$. Putting these facts together, we obtain

$$\eta_3 \leq O(p^{3/2}n^{-1}\sqrt{n^{-1}\log p}) = o(\sqrt{n^{-1}\log n}) \qquad (40)$$

where we use $p = O(n^{\alpha_0})$ with $\alpha_0 \in [0, 2/3)$.

Combining equations (35), (39), and (40), we obtain that with probability at least $1 - o(p^{-a})$,

$$\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu}_0)\|_\infty \leq c\sqrt{n^{-1}\log n}.$$

■

Lemma 6 *Under the assumptions of Theorem 4,* $\|(n^{-1}\mathbf{X}^T\mathbf{X})^{-1}\|_\infty \leq 1 + O(pn^{-1/2})$ *with probability* $1 - O(p^{-c})$.

Proof Let $\mathbf{E} = n^{-1}\mathbf{X}^T\mathbf{X} - \mathbf{I}_p$. Then, $\|\mathbf{E}\|_2 \leq C(p/n)^{1/2}$ for some constant C with probability $1 - O(p^{-c})$ by Theorem 4.6.1 in Vershynin (2016). Furthermore, by matrix inversion, we get

$$(n^{-1}\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{I}_p + \mathbf{E})^{-1} = \mathbf{I}_p - \sum_{k=1}^{\infty}(-1)^{k+1}(\mathbf{E})^k.$$

Now, we take the norm and use triangle inequalities to get

$$\begin{aligned}\|(n^{-1}\mathbf{X}^T\mathbf{X})^{-1}\|_\infty &\leq \|\mathbf{I}_p\|_\infty + \sum_{k=1}^{\infty}\|\mathbf{E}^k\|_\infty \leq 1 + p^{1/2}\sum_{k=1}^{\infty}\|\mathbf{E}^k\|_2 \\ &\leq 1 + p^{1/2}\sum_{k=1}^{\infty}\|\mathbf{E}\|_2^k \leq 1 + Cp^{1/2}\sum_{k=1}^{\infty}((p/n)^{1/2})^k \\ &\leq 1 + Cp^{1/2}(p/n)^{1/2}\end{aligned}$$

where we use the fact that p/n is bounded by a constant less than 1. ■

Lemma 7 *In the same setting as Lemma 6, if* $\mathbf{E} = n^{-1}\mathbf{X}^T\mathbf{X} - \mathbf{I}_p$, *then* $\left\|\sum_{k=2}^{\infty}(-1)^{k+1}(\mathbf{E})^k\right\|_\infty \leq Cp^{3/2}n^{-1}$, *with probability* $1 - O(p^{-c})$.

Proof Again, we use that $\|\mathbf{E}\|_2 \leq C(p/n)^{1/2}$ for some constant C with probability $1 - O(p^{-c})$. By similar calculations as in Lemma 6, we deduce

$$\begin{aligned}\left\|\sum_{k=2}^{\infty}(-1)^{k+1}(\mathbf{E})^k\right\|_\infty &\leq \sum_{k=2}^{\infty}\left\|(-1)^{k+1}(\mathbf{E})^k\right\|_\infty \leq \sum_{k=2}^{\infty}p^{1/2}\left\|(\mathbf{E})^k\right\|_2 \\ &= \sum_{k=2}^{\infty}p^{1/2}\|\mathbf{E}\|_2^k \leq \sum_{k=2}^{\infty}p^{1/2}(p/n)^{k/2} \leq Cp^{3/2}n^{-1}.\end{aligned}$$

■

Lemma 8 *Let* $\mathbf{W}_j$ *be nonnegative random variables for* $1 \leq j \leq p$ *that are not necessarily independent. If* $P(W_j > t) \leq C_1\exp(-C_2a_nt^2)$ *for some constants* C_1 *and* C_2 *and for some sequence* a_n, *then for any* $c > 0$, $\max_{1\leq j\leq p}W_j \leq ((c+1)/C_2)^{1/2}a_n^{-1/2}(\log p)^{1/2}$ *with probability at least* $1 - O(p^{-c})$.

Proof Using union bound, we get

$$P(\max_{1\leq j\leq p}W_j > t) \leq \sum_{1\leq j\leq p}P(W_j > t) \leq pC_1\exp(-C_2a_nt^2).$$

Taking $t = a_n^{-1/2}(\log p)^{1/2}((c+1)/C_2)^{1/2}$ concludes the proof since then

$$P(\max_{1\leq j\leq p}W_j > a_n^{-1/2}(\log p)^{1/2}((c+1)/C_2)^{1/2}) \leq C_1p^{-c}.$$

■

Lemma 9 *Let W_{ij} be random variables which are independent over the index i. Assume that there are constants C_1 and C_2 such that $P(|W_{ij}| > t) \leq C_1 \exp(-C_2 t^\alpha)$ with $0 < \alpha \leq 1$. Then, with probability $1 - O(p^{-c})$,*

$$\max_{0 \leq j \leq p} \left| n^{-1} \sum_{i=1}^{n} (W_{ij} - EW_{ij}) \right| \leq C n^{-(1/2)\alpha} (\log p)^{1/2},$$

for some positive constants c and C.

Proof We have $P(|n^{-1} \sum_{i=1}^{n} (W_{ij} - EW_{ij})| > t) \leq C_3 \exp(-C_4 n^\alpha t^2)$ by Lemma 6 of Fan et al. (2016) where C_3 and C_4 are some positive constants which only depend on C_1 and C_2. This probability bound shows that the assumption of Lemma 8 holds with $a_n = n^\alpha$. Thus, Lemma 8 finishes the proof. ■

Lemma 10 *With probability $1 - O(p^{-c})$, the vector $\mathbf{r}$ defined in (27) satisfies the bound $\|n^{-1}\mathbf{r}\|_\infty = O(n^{\alpha - 5/6} \log n \sqrt{n^{-1} \log n})$.*

Proof We begin by observing that both x_{ij} and $(\mathbf{x}_i^T \boldsymbol{\beta} / \|\boldsymbol{\beta}\|_2)$ are standard normal variables. So, using Lemma 1 of Fan et al. (2016), we have $P(x_{ij}(\mathbf{x}_i^T \boldsymbol{\beta}/\|\boldsymbol{\beta}\|_2)^2 > t) \leq C \exp(-Ct^{2/3})$ for some constant C which does not depend $\boldsymbol{\beta}$. It is easy to see that $x_{ij}(\mathbf{x}_i^T \boldsymbol{\beta})^2$ are independent random variables across i's with mean 0. By Lemma 9, $\max_{1 \leq j \leq p} |n^{-1} \sum_{i=1}^{n} x_{ij} \left(\frac{\mathbf{x}_i^T \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2} \right)^2|$ is of order $O(n^{-1/3} (\log p)^{1/2})$. Moreover, $\|\boldsymbol{\beta}\|_2 \leq p^{1/2} \|\boldsymbol{\beta}\|_\infty \leq O(p^{1/2} \sqrt{n^{-1} \log n})$ when $\boldsymbol{\beta} \in \mathcal{N}_0$. Therefore,

$$\begin{aligned}
\|n^{-1}\mathbf{r}\|_\infty &= \max_{1 \leq j \leq p} \left| \frac{b^{(3)}(0)}{2} \|\boldsymbol{\beta}\|_2^2 n^{-1} \sum_{i=1}^{n} x_{ij} (\mathbf{x}_i^T \boldsymbol{\beta}/\|\boldsymbol{\beta}\|_2)^2 \right| \\
&\leq Cpn^{-1} (\log n) n^{-1/3} (\log p)^{1/2} = Cn^{\alpha - 4/3} (\log n)^{3/2} \\
&= O(n^{\alpha - 5/6} (\log n) \sqrt{n^{-1} \log n}),
\end{aligned}$$

since $p = O(n^\alpha)$. ■

Lemma 11 *With probability $1 - O(p^{-c})$, the vector $\mathbf{s}$ defined in (28) satisfies the bound $\|n^{-1}\mathbf{s}\|_\infty = O((n^{3/2\alpha - 5/4} (\log n)^{3/2} + (n^{\alpha - 1} (\log n)^{1/2}) \sqrt{n^{-1} \log n})$.*

Proof First, observe that for some constant C, $|n^{-1} s_j| \leq C \|\boldsymbol{\beta}\|_2^3 n^{-1} \sum_{i=1}^{n} x_{ij} \left(\frac{\mathbf{x}_i^T \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2} \right)^3$. Moreover, the summands $x_{ij} \left(\frac{\mathbf{x}_i^T \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2} \right)^3$ are independent over i and they satisfy the probability bound $P(|x_{ij} \left(\frac{\mathbf{x}_i^T \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2} \right)^3| > t) \leq C \exp(-Ct^{1/2})$ by Lemma 1 of Fan et al. (2016). Thus, by Lemma 9, we obtain

$$\max_{1 \leq j \leq p} \left| n^{-1} \sum_{i=1}^{n} \left(x_{ij} \left(\frac{\mathbf{x}_i^T \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2} \right)^3 - E \left[\left(x_{ij} \frac{\mathbf{x}_i^T \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2} \right)^3 \right] \right) \right| = O(n^{-1/4} (\log p)^{1/2}).$$

Now, we calculate the expected value of the summand $x_{ij}\left(\frac{\mathbf{x}_i^T\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2}\right)^3$. We decompose $\mathbf{x}_i^T\boldsymbol{\beta}$ as $x_{ij}\beta_j + \mathbf{x}_{i,-j}^T\boldsymbol{\beta}_{-j}$ where $\mathbf{x}_{i,-j}$ and $\boldsymbol{\beta}_{-j}$ are the vectors $\mathbf{x}_i$ and $\boldsymbol{\beta}$ whose jth entry is removed. We use the independence of $\mathbf{x}_{i,-j}$ and x_{ij} and get

$$\begin{aligned}
E\left[x_{ij}\left(\frac{\mathbf{x}_i^T\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2}\right)^3\right] &= \frac{1}{\|\boldsymbol{\beta}\|_2^3}E\left[x_{ij}\left(x_{ij}\beta_j + \mathbf{x}_{i,-j}^T\boldsymbol{\beta}_{-j}\right)^3\right] \\
&= \frac{1}{\|\boldsymbol{\beta}\|_2^3}E\left[x_{ij}^4\beta_j^3 + 3x_{ij}^3\beta_j^2\left(\mathbf{x}_{i,-j}^T\boldsymbol{\beta}_{-j}\right) + 3x_{ij}^2\beta_j\left(\mathbf{x}_{i,-j}^T\boldsymbol{\beta}_{-j}\right)^2 + x_{ij}\left(\mathbf{x}_{i,-j}^T\boldsymbol{\beta}_{-j}\right)^3\right] \\
&= \frac{1}{\|\boldsymbol{\beta}\|_2^3}\left[3\beta_j^3 + 3\beta_j\|\boldsymbol{\beta}_{-j}\|_2^2\right] \\
&= \frac{3\beta_j}{\|\boldsymbol{\beta}\|_2}.
\end{aligned}$$

Finally, we can combine the result of Lemma 9 and the expected value of $x_{ij}\left(\frac{\mathbf{x}_i^T\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2}\right)^3$. We bound $\|n^{-1}\mathbf{s}\|_\infty$ as follows

$$\begin{aligned}
\|n^{-1}\mathbf{s}\|_\infty &\le C\|\boldsymbol{\beta}\|_2^3 \max_{1\le j\le p}\left|n^{-1}\sum_{i=1}^n x_{ij}\left(\frac{\mathbf{x}_i^T\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2}\right)^3\right| \\
&\le O\left(\|\boldsymbol{\beta}\|_2^3(n^{-1/4}(\log p)^{1/2} + \frac{\|\boldsymbol{\beta}\|_\infty}{\|\boldsymbol{\beta}\|_2})\right) \\
&\le O\left(\|\boldsymbol{\beta}\|_2^3 n^{-1/4}(\log p)^{1/2} + \|\boldsymbol{\beta}\|_\infty\|\boldsymbol{\beta}\|_2^2)\right).
\end{aligned}$$

Since $\boldsymbol{\beta}\in\mathcal{N}_0$, we have $\|\boldsymbol{\beta}\|_2 = O(p^{1/2}n^{-1/2}(\log p)^{1/2})$ and $\|\boldsymbol{\beta}\|_\infty = O(n^{-1/2}(\log p)^{1/2})$. Thus, $\|n^{-1}\mathbf{s}\|_\infty = O((n^{3/2\alpha-5/4}(\log n)^{3/2} + (n^{\alpha-1}(\log n)^{1/2})\sqrt{n^{-1}\log n})$ when $p = O(n^\alpha)$. ■

Lemma 12 *With probability* $1 - O(p^{-c})$*, the vector* $\mathbf{t}$ *defined in* (29) *satisfies the bound* $\|n^{-1}\mathbf{t}\|_\infty = O(n^{2\alpha-3/2}(\log n)^{3/2}\sqrt{n^{-1}\log n})$.

Proof The proof is similar to the proof of Lemma 11. Since $b^{(5)}(\cdot)$ is uniformly bounded, $|n^{-1}t_j| \le C\|\boldsymbol{\beta}\|_2^4 n^{-1}\sum_{i=1}^n\left|x_{ij}\left(\frac{\mathbf{x}_i^T\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2}\right)^4\right|$ for some constant C. We focus on the summands $\left|x_{ij}\left(\frac{\mathbf{x}_i^T\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2}\right)^4\right|$ which are independent across i. Moreover, repeated application of Lemma 1 of Fan et al. (2016) yields $P(x_{ij}(\mathbf{x}_i^T\boldsymbol{\beta}/\|\boldsymbol{\beta}\|_2)^2 > t) \le C\exp(-Ct^{2/5})$ for some constant C independent of $\boldsymbol{\beta}$. We can bound the expected value of the summand by Cauchy-Schwartz: $E\left[\left|x_{ij}\left(\frac{\mathbf{x}_i^T\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2}\right)^4\right|\right] \le \left(Ex_{ij}^2E\left[\left(\frac{\mathbf{x}_i^T\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2}\right)^8\right]\right)^{1/2} = \sqrt{105}$. So, by Lemma 9, we get

$$\begin{aligned}
\|n^{-1}\mathbf{t}\|_\infty &\le C\|\boldsymbol{\beta}\|_2^4(\sqrt{105} + n^{-1/5}(\log p)^{1/2}) \\
&= O(\|\boldsymbol{\beta}\|_2^4) = O(p^2n^{-2}(\log n)^2).
\end{aligned}$$

Finally, we can deduce that $\|n^{-1}\mathbf{t}\|_\infty = O(n^{2\alpha-3/2}(\log n)^{3/2}\sqrt{n^{-1}\log n})$ when $p = O(n^\alpha)$. ■

Lemma 13 *Let* $T = \mathbf{e}_j^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}_0] - n^{-1}\mathbf{e}_j^T\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}_0]$. *Under the assumptions of Theorem 4, we have*

$$\mathbf{e}_j^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}_0] - n^{-1}\mathbf{e}_j^T\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}_0] = o_p(n^{-1/2}). \tag{41}$$

Proof Since $\mathbf{X}$ and $\mathbf{y}$ are independent, expectation of T is clearly zero. Then, we consider the variance of T. To this end, we condition on X. We can calculate the conditional variance of T as follows

$$\begin{aligned}\mathrm{var}[T|\mathbf{X}] &= \mathrm{var}[\mathbf{e}_j^T((\mathbf{X}^T\mathbf{X})^{-1} - n^{-1}I_p)\mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu}_0)|\mathbf{X}] \\ &= \phi b''(0)\mathbf{e}_j^T((\mathbf{X}^T\mathbf{X})^{-1} - n^{-1}I_p)\mathbf{X}^T\mathbf{X}((\mathbf{X}^T\mathbf{X})^{-1} - n^{-1}I_p)\mathbf{e}_j\end{aligned}$$

where we use $\mathrm{var}[y] = \phi b''(0)I_n$. When we define $\mathbf{E} = n^{-1}\mathbf{X}^T\mathbf{X} - I_p$, simple calculations show that

$$\begin{aligned}\mathrm{Var}[T|\mathbf{X}] &= \phi b''(0)n^{-1}\mathbf{e}_j^T((n^{-1}\mathbf{X}^T\mathbf{X})^{-1} - I_p) + (n^{-1}\mathbf{X}^T\mathbf{X} - I_p))\mathbf{e}_j \\ &= \phi b''(0)n^{-1}\mathbf{e}_j^T\left(\sum_{k=2}^{\infty}(-1)^k\mathbf{E}^k\right)\mathbf{e}_j.\end{aligned}$$

Now, we can obtain the unconditional variance using the law of total variance.

$$\begin{aligned}\mathrm{var}[T] &= E[\mathrm{var}[T|\mathbf{X}]] + \mathrm{var}[E[T|\mathbf{X}]] \\ &= \phi b''(0)n^{-1}\mathbf{e}_j^T E\left(\sum_{k=2}^{\infty}(-1)^k\mathbf{E}^k\right)\mathbf{e}_j.\end{aligned}$$

Thus, using Lemma 7, we can show that $\mathrm{var}[T] = o(n^{-1})$. Finally, we use Chebyshev's inequality $P(|T| > n^{-1/2}) \le n\mathrm{var}[T] = o(1)$. So, we conclude that $T = o_p(n^{-1/2})$ ■

Lemma 14 *Let* x_{ij} *be standard normal random variables for* $1 \le i \le n$ *and* $1 \le j \le p$. *Then,* $\max_{1\le j\le p}\sum_{i=1}^n x_{ij}^2 \le n + O(n^{1/2}(\log p)^{1/2})$ *with probability* $1 - O(p^{-c})$ *for some positive constant c. Consequently, when* $\log p = O(n^\alpha)$ *for some* $0 < \alpha \le 1$, *we have* $\max_{1\le j\le p}\sum_{i=1}^n x_{ij}^2 = O(n)$, *for large enough n with probability* $1 - O(p^{-c})$.

Proof Since x_{ij} is a standard normal variable, x_{ij}^2 is subexponential random variable whose mean is 1. So, Lemma 9 entails that

$$\max_{1\le j\le p}\left|n^{-1}\sum_{i=1}^n(x_{ij}^2 - 1)\right| = O(n^{-1/2}(\log p)^{1/2})$$

with probability $1 - O(p^{-c})$. Thus, simple calculations yields

$$\max_{1\le j\le p}\sum_{i=1}^n x_{ij}^2 = \max_{1\le j\le p}\left|n + \sum_{i=1}^n(x_{ij}^2 - 1)\right| \le n + O(n^{1/2}(\log p)^{1/2})$$

with probability $1 - O(p^{-c})$. ■

B.3. Proof of Theorem 2

To prove the conclusion in Theorem 2, we use the proof by contradiction. Let us make an assumption (A) that the asymptotic normality (11) in Theorem 1 which has been proved to hold when $p = o(n^{1/2})$ continues to hold when $p \sim n^{\alpha_0}$ for some constant $1/2 < \alpha_0 \leq 1$, where $\sim$ stands for asymptotic order. As shown in Section 3.1, in the case of logistic regression under global null (that is, $\boldsymbol{\beta}_0 = \mathbf{0}$) with deterministic rescaled orthonormal design matrix $\mathbf{X}$ (in the sense of $n^{-1}\mathbf{X}^T\mathbf{X} = I_p$) the limiting distribution in (11) by assumption (A) becomes

$$2^{-1}n^{1/2}\widehat{\beta}_j \xrightarrow{\mathscr{D}} N(0,1), \tag{42}$$

where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T$ is the MLE.

Let us now assume that the rescaled random design matrix $n^{-1/2}\mathbf{X}$ is uniformly distributed on the Stiefel manifold $V_p(\mathbb{R}^n)$ which can be thought of as the space of all $n \times p$ orthonormal matrices. Then it follows from (42) that

$$2^{-1}n^{1/2}\widehat{\beta}_j \xrightarrow{\mathscr{D}} N(0,1) \ \text{ conditional on } \mathbf{X}. \tag{43}$$

Based on the limiting distribution in (43), we can make two observations. First, it holds that

$$2^{-1}n^{1/2}\widehat{\beta}_j \xrightarrow{\mathscr{D}} N(0,1) \tag{44}$$

unconditional on the design matrix $\mathbf{X}$. Second, $\widehat{\beta}_j$ is asymptotically independent of the design matrix $\mathbf{X}$, and so is the MLE $\widehat{\boldsymbol{\beta}}$.

Since the distribution of $n^{-1/2}\mathbf{X}$ is assumed to be the Haar measure on the Stiefel manifold $V_p(\mathbb{R}^n)$, we have

$$n^{-1/2}\mathbf{X}\mathbf{Q} \overset{d}{=} n^{-1/2}\mathbf{X}, \tag{45}$$

where $\mathbf{Q}$ is any fixed $p \times p$ orthogonal matrix and $\overset{d}{=}$ stands for equal in distribution. Recall that the MLE $\widehat{\boldsymbol{\beta}}$ solves the score equation (4), which is in turn equivalent to equation

$$\mathbf{Q}^T\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})] = \mathbf{0} \tag{46}$$

since $\mathbf{Q}$ is orthogonal. We now use the fact that the model is under global null which entails that the response vector $\mathbf{y}$ is independent of the design matrix $\mathbf{X}$. Combining this fact with (45)–(46) yields

$$\mathbf{Q}^T\widehat{\boldsymbol{\beta}} \overset{d}{=} \widehat{\boldsymbol{\beta}} \tag{47}$$

by noting that $\mathbf{X}\boldsymbol{\beta} = (\mathbf{X}\mathbf{Q})(\mathbf{Q}^T\boldsymbol{\beta})$. Since the distributional identity (47) holds for any fixed $p \times p$ orthogonal matrix $\mathbf{Q}$, we conclude that the MLE $\widehat{\boldsymbol{\beta}}$ has a spherical distribution on $\mathbb{R}^p$. It is a well-known fact that all the marginal characteristic functions of a spherical distribution have the same generator. Such a fact along with (44) entails that

$$2^{-1}n^{1/2}\widehat{\boldsymbol{\beta}} \text{ is asymptotically close to } N(\mathbf{0}, I_p). \tag{48}$$

To simplify the exposition, let us now make the asymptotic limit exact and assume that

$$\widehat{\boldsymbol{\beta}} \sim N(\mathbf{0}, 4n^{-1}I_p) \text{ and is independent of } \mathbf{X}. \tag{49}$$

The remaining analysis focuses on the score equation (4) which is solved exactly by the MLE $\widehat{\boldsymbol{\beta}}$, that is,

$$\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}})] = \mathbf{0}, \tag{50}$$

which leads to

$$\boldsymbol{\xi} \equiv n^{-1/2}\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{0})] = n^{-1/2}\mathbf{X}^T[\boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}) - \boldsymbol{\mu}(\mathbf{0})] \equiv \boldsymbol{\eta}. \tag{51}$$

Let us first consider the random variable $\boldsymbol{\xi}$ defined in (51). Note that $2[\mathbf{y} - \boldsymbol{\mu}(\mathbf{0})]$ has independent and identically distributed (i.i.d.) components each taking value 1 or -1 with equal probability $1/2$, and is independent of $\mathbf{X}$. Thus since $n^{-1/2}\mathbf{X}$ is uniformly distributed on the Stiefel manifold $V_p(\mathbb{R}^n)$, it is easy to see that

$$\boldsymbol{\xi} = n^{-1/2}\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{0})] \stackrel{d}{=} 2^{-1}n^{-1/2}\mathbf{X}^T\mathbf{1}, \tag{52}$$

where $\mathbf{1} \in \mathbb{R}^n$ is a vector with all components being one. Using similar arguments as before, we can show that $\boldsymbol{\xi}$ has a spherical distribution on $\mathbb{R}^p$. Thus the joint distribution of $\boldsymbol{\xi}$ is determined completely by the marginal distribution of $\boldsymbol{\xi}$. For each $1 \le j \le p$, denote by ξ_j the jth component of $\boldsymbol{\xi} = 2^{-1}n^{-1/2}\mathbf{X}^T\mathbf{1}$ using the distributional representation in (52). Let $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ with each $\mathbf{x}_j \in \mathbb{R}^n$. Then we have

$$\xi_j = 2^{-1}n^{-1/2}\mathbf{x}_j^T\mathbf{1} \stackrel{d}{=} 2^{-1}(n^{1/2}/\|\widetilde{\mathbf{x}}_j\|_2)n^{-1/2}\widetilde{\mathbf{x}}_j^T\mathbf{1}, \tag{53}$$

where $\widetilde{\mathbf{x}}_j \sim N(\mathbf{0}, 4^{-1}I_n)$. It follows from (53) and the concentration phenomenon of Gaussian measures that each ξ_j is asymptotically close to $N(0, 4^{-1})$ and thus consequently $\boldsymbol{\xi}$ is asymptotically close to $N(\mathbf{0}, 4^{-1}I_p)$. *A key fact (i) for the finite-sample distribution of $\boldsymbol{\xi}$ is that the standard deviation of each component ξ_j converges to $1/2$ at rate $O_P(n^{-1/2})$ that does not depend upon the dimensionality p at all.*

We now turn our attention to the second term $\boldsymbol{\eta}$ defined in (51). In view of (49) and the fact that $n^{-1/2}\mathbf{X}$ is uniformly distributed on the Stiefel manifold $V_p(\mathbb{R}^n)$, we can show that with significant probability,

$$\|\mathbf{X}\widehat{\boldsymbol{\beta}}\|_\infty \le o(1) \tag{54}$$

for $p \sim n^{\alpha_0}$ with $\alpha_0 < 1$. The uniform bound in (54) enables us to apply the mean value theorem for the vector-valued function $\boldsymbol{\eta}$ around $\boldsymbol{\beta}_0 = \mathbf{0}$, which results in

$$\begin{aligned} \boldsymbol{\eta} &= n^{-1/2}\mathbf{X}^T[\boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}) - \boldsymbol{\mu}(\mathbf{0})] = 4^{-1}n^{-1/2}\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{r} \\ &= 4^{-1}n^{1/2}\widehat{\boldsymbol{\beta}} + \mathbf{r} \end{aligned} \tag{55}$$

since $n^{-1/2}\mathbf{X}$ is assumed to be orthonormal, where

$$\mathbf{r} = n^{-1/2}\mathbf{X}^T\left\{\int_0^1 \left[\boldsymbol{\Sigma}(t\mathbf{X}\widehat{\boldsymbol{\beta}}) - 4^{-1}I_n\right] dt\right\}\mathbf{X}\widehat{\boldsymbol{\beta}}. \tag{56}$$

Here, the remainder term $\mathbf{r} = (r_1, \cdots, r_p)^T \in \mathbb{R}^p$ is stochastic and each component r_j is generally of order $O_P\{p^{1/2}n^{-1/2}\}$ in light of (49) when the true model may deviate from the global null case of $\boldsymbol{\beta}_0 = \mathbf{0}$.

Since our focus in this theorem is the logistic regression model under the global null, we can in fact claim that each component r_j is generally of order $O_P\{pn^{-1}\}$, which is a better rate of convergence than the one mentioned above thanks to the assumption of $\boldsymbol{\beta}_0 = \mathbf{0}$. To prove this claim, note that the variance function $b''(\theta)$ is symmetric in $\theta \in \mathbb{R}$ and takes the maximum value $1/4$ at $\theta = 0$. Thus in view of (54), we can show that with significant probability,

$$4^{-1}I_n - \boldsymbol{\Sigma}(t\mathbf{X}\widehat{\boldsymbol{\beta}}) \geq c\mathrm{diag}\{(t\mathbf{X}\widehat{\boldsymbol{\beta}}) \circ (t\mathbf{X}\widehat{\boldsymbol{\beta}})\} = ct^2\mathrm{diag}\{(\mathbf{X}\widehat{\boldsymbol{\beta}}) \circ (\mathbf{X}\widehat{\boldsymbol{\beta}})\} \tag{57}$$

for all $t \in [0,1]$, where $c > 0$ is some constant and $\geq$ stands for the inequality for positive semidefinite matrices. Moreover, it follows from (49) and the fact that $n^{-1/2}\mathbf{X}$ is uniformly distributed on the Stiefel manifold $V_p(\mathbb{R}^n)$ that with significant probability, all the n components of $\mathbf{X}\widehat{\boldsymbol{\beta}}$ are concentrated in the order of $p^{1/2}n^{-1/2}$. This result along with (57) and the fact that $n^{-1}\mathbf{X}^T\mathbf{X} = I_p$ entails that with significant probability,

$$\begin{aligned} & n^{-1/2}\mathbf{X}^T\left\{\int_0^1 \left[4^{-1}I_n - \boldsymbol{\Sigma}(t\mathbf{X}\widehat{\boldsymbol{\beta}})\right]dt\right\}\mathbf{X} \\ & \geq n^{-1/2}\mathbf{X}^T\left\{\int_0^1 c_* t^2 pn^{-1}dt\right\}\mathbf{X} \\ & = 3^{-1}c_* pn^{-3/2}\mathbf{X}^T\mathbf{X} = 3^{-1}c_* pn^{-1/2}I_p, \end{aligned} \tag{58}$$

where $c_* > 0$ is some constant. Thus combining (56), (58), and (49) proves the above claim.

We make two important observations about the remainder term $\mathbf{r}$ in (55). First, $\mathbf{r}$ has a spherical distribution on $\mathbb{R}^p$. This is because by (55) and (51) it holds that

$$\mathbf{r} = \boldsymbol{\eta} - 4^{-1}n^{1/2}\widehat{\boldsymbol{\beta}} = \boldsymbol{\xi} - 4^{-1}n^{1/2}\widehat{\boldsymbol{\beta}},$$

which has a spherical distribution on $\mathbb{R}^p$. Thus the joint distribution of $\mathbf{r}$ is determined completely by the marginal distribution of $\mathbf{r}$. Second, for the nonlinear setting of logistic regression model, the appearance of the remainder term $\mathbf{r}$ in (55) is due solely to *the nonlinearity of the mean function* $\boldsymbol{\mu}(\cdot)$, and we have shown that each component r_j can indeed achieve the worst-case order pn^{-1} in probability. For each $1 \leq j \leq p$, denote by η_j the jth component of $\boldsymbol{\eta}$. Then in view of (49) and (55), *a key fact (ii) for the finite-sample distribution of* $\boldsymbol{\eta}$ *is that the standard deviation of each component* η_j *converges to* $1/2$ *at rate* $O_P\{pn^{-1}\}$ *that generally does depend upon the dimensionality* p.

Finally, we are ready to compare the two random variables $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ on the two sides of equation (51). Since equation (51) is a distributional identity in $\mathbb{R}^p$, naturally the square root of the sum of $\mathrm{var}\xi_j$'s and the square root of the sum of $\mathrm{var}\eta_j$'s are expected to converge to the common value $2^{-1}p^{1/2}$ at rates that are asymptotically negligible. However, the former has rate $p^{1/2}O_P(n^{-1/2}) = O_P\{p^{1/2}n^{-1/2}\}$, whereas the latter has rate $p^{1/2}O_P\{pn^{-1}\} = O_P\{p^{3/2}n^{-1}\}$. A key consequence is that when $p \sim n^{\alpha_0}$ for some constant $2/3 \leq \alpha_0 < 1$, there is a profound difference between the two asymptotic rates in that the former rate is $O_P\{n^{-(1-\alpha_0)/2}\} = o_P(1)$, while the latter rate becomes $O_P\{n^{3\alpha_0/2-1}\}$ which is now asymptotically diverging or nonvanishing. Such an intrinsic asymptotic difference is, however, prohibited by the distributional identity (51) in $\mathbb{R}^p$, which results in a contradiction. Therefore, we have now argued that assumption (A) we started with for

$2/3 \leq \alpha_0 < 1$ must be false, that is, the asymptotic normality (11) which has been proved to hold when $p = o(n^{1/2})$ generally would not continue to hold when $p \sim n^{\alpha_0}$ with constant $2/3 \leq \alpha_0 \leq 1$. In other words, we have proved the invalidity of the conventional GLM p-values in this regime of diverging dimensionality, which concludes the proof of Theorem 2.

B.4. Proof of Theorem 3

By assumption, $\mathbf{X} \sim N(\mathbf{0}, I_n \otimes \mathbf{\Sigma})$ with covariance matrix $\mathbf{\Sigma}$ nonsingular. Let us first make a useful observation. For the general case of nonsingular covariance matrix $\mathbf{\Sigma}$, we can introduce a change of variable by letting $\widetilde{\boldsymbol{\beta}} = \mathbf{\Sigma}^{1/2}\boldsymbol{\beta}$ and correspondingly $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{\Sigma}^{-1/2}$. Clearly, $\widetilde{\mathbf{X}} \sim N(\mathbf{0}, I_n \otimes I_p)$ and the MLE for the transformed parameter vector $\widetilde{\boldsymbol{\beta}}$ is exactly $\mathbf{\Sigma}^{1/2}\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ denotes the MLE under the original design matrix $\mathbf{X}$. Thus to show the breakdown point of the conventional asymptotic normality of the MLE, it suffices to focus on the specific case of $\mathbf{X} \sim N(\mathbf{0}, I_n \otimes I_p)$.

Hereafter we assume that $\mathbf{X} \sim N(\mathbf{0}, I_n \otimes I_p)$ with $p = o(n)$. The rest of the arguments are similar to those in the proof of Theorem 2 in Section B.3 except for some modifications needed for the case of Gaussian design. Specifically, for the case of logistic regression model under global null (that is, $\boldsymbol{\beta}_0 = \mathbf{0}$), the limiting distribution in (11) becomes

$$2^{-1}n^{1/2}\widehat{\beta}_j \xrightarrow{\mathscr{D}} N(0, 1), \tag{59}$$

since $n^{-1}\mathbf{X}^T\mathbf{X} \to I_p$ almost surely in spectrum and thus $4^{-1}n(\mathbf{A}_n^{-1})_{jj} \to 1$ in probability as $n \to \infty$. Here, we have used a claim that both the largest and smallest eigenvalues of $n^{-1}\mathbf{X}^T\mathbf{X}$ converge to 1 almost surely as $n \to \infty$ for the case of $p = o(n)$, which can be shown by using the classical results from random matrix theory (RMT) Geman (1980); Silverstein (1985); Bai (1999).

Note that since $\mathbf{X} \sim N(0, I_n \otimes I_p)$, it holds that

$$n^{-1/2}\mathbf{X}\mathbf{Q} \overset{d}{=} n^{-1/2}\mathbf{X}, \tag{60}$$

where $\mathbf{Q}$ is any fixed $p \times p$ orthogonal matrix and $\overset{d}{=}$ stands for equal in distribution. By $\mathbf{X} \sim N(0, I_n \otimes I_p)$, it is also easy to see that

$$\boldsymbol{\xi} = n^{-1/2}\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{0})] \overset{d}{=} 2^{-1}n^{-1/2}\mathbf{X}^T\mathbf{1}, \tag{61}$$

where $\mathbf{1} \in \mathbb{R}^n$ is a vector with all components being one. In view of (49) and the assumption of $\mathbf{X} \sim N(0, I_n \otimes I_p)$, we can show that with significant probability,

$$\|\mathbf{X}\widehat{\boldsymbol{\beta}}\|_\infty \leq o(1) \tag{62}$$

for $p \sim n^{\alpha_0}$ with constant $\alpha_0 < 1$. It holds further that with significant probability, all the n components of $\mathbf{X}\widehat{\boldsymbol{\beta}}$ are concentrated in the order of $p^{1/2}n^{-1/2}$. This result along with (57) and the fact that $n^{-1}\mathbf{X}^T\mathbf{X} \to I_p$ almost surely in spectrum entails that with asymptotic

probability one,

$$
\begin{aligned}
& n^{-1/2}\mathbf{X}^T\left\{\int_0^1\left[4^{-1}I_n-\boldsymbol{\Sigma}(t\mathbf{X}\widehat{\boldsymbol{\beta}})\right]dt\right\}\mathbf{X} \qquad (63)\\
&\geq n^{-1/2}\mathbf{X}^T\left\{\int_0^1 c_*t^2pn^{-1}dt\right\}\mathbf{X}\\
&= 3^{-1}c_*pn^{-3/2}\mathbf{X}^T\mathbf{X}\to 3^{-1}c_*pn^{-1/2}I_p,
\end{aligned}
$$

where $c_*>0$ is some constant. This completes the proof of Theorem 3.

References

T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23:193–212, 1952.

T. W. Anderson and D. A. Darling. A test of goodness-of-fit. *Journal of the American Statistical Association*, 49:765–769, 1954.

Susan Athey, Guido W. Imbens, and Stefan Wager. Efficient inference of average treatment effects in high dimensions via approximate residual balancing. *arXiv preprint arXiv:1604.07125*, 2016.

Z. D. Bai. Methodologies in spectral analysis of large dimensional random matrices, a review. *Statist. Sin.*, 9:611–677, 1999.

Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43:2055–2085, 2015.

Derek Bean, Peter J. Bickel, Noureddine E. Karoui, and Bin Yu. Optimal M-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences of the United States of America*, 110:14563–14568, 2013.

E. J. Candès. Private communication. 2016.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B*, 80:551–577, 2018.

J. Fan and H. Peng. Nonconcave penalized likelihood with diverging number of parameters. *Ann. Statist.*, 32:928–961, 2004.

Jianqing Fan and Jinchi Lv. Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, 57:5467–5484, 2011.

Y. Fan and J. Lv. Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association*, 108:1044–1061, 2013.

Yingying Fan, Yinfei Kong, Daoji Li, and Jinchi Lv. Interaction pursuit with feature screening and selection. *arXiv preprint arXiv:1605.08933*, 2016.

Yingying Fan, Emre Demirkaya, Gaorong Li, and Jinchi Lv. RANK: large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association, to appear*, 2018.

S. Geman. A limit theorem for the norm of random matrices. *Ann. Probab.*, 8:252–261, 1980.

Bin Guo and Song Xi Chen. Tests for high dimensional generalized linear models. *J. R. Statist. Soc. B*, 78:1079–1102, 2016.

P. J. Huber. Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1:799–821, 1973.

Noureddine E. Karoui, Derek Bean, Peter J. Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences of the United States of America*, 110:14557–14562, 2013.

A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari.*, 4:83–91, 1933.

J. Lv and J. S. Liu. Model selection principles in misspecified models. *Journal of the Royal Statistical Society Series B*, 76:141–167, 2014.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989.

S. Portnoy. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. i. consistency. *The Annals of Statistics*, 12:1298–1309, 1984.

S. Portnoy. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; ii. normal approximation. *The Annals of Statistics*, 13:1403–1417, 1985.

S. Portnoy. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, 16:356–366, 1988.

J. W. Silverstein. The smallest eigenvalue of a large dimensional wishart matrix. *Ann. Probab.*, 13:1364–1368, 1985.

N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19:279–281, 1948.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *arXiv preprint arXiv:1803.06964*, 2018.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42:1166–1202, 2014.

Roman Vershynin. High-dimensional probability. *An Introduction with Applications*, 2016.

Michael N Vrahatis. A short proof and a generalization of Miranda's existence theorem. *Proceedings of the American Mathematical Society*, 107:701–703, 1989.